DOCUMENT RESUME

ED 332 751 JC 910 270

AUTHOR Rubadeau, Duane O.; And Others
TITLE Appropriate Testing
INSTITUTION College of New Caledonia, Prince George (British Columbia). Centre for Improved Teaching.
REPORT NO ISBN-0-921087-16-0
PUB DATE 90
NOTE 101p.
PUB TYPE Guides - Classroom Use - Teaching Guides (For Teacher) (052)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Achievement Tests; Community Colleges; *Educational Testing; Elementary Secondary Education; Foreign Countries; *Standardized Tests; *Teacher Made Tests; *Test Construction; Test Format; *Testing Problems; Test Selection; *Test Use; Two Year Colleges

ABSTRACT

Intended as a guide in the development of a sound evaluation program utilizing both instructor-produced and standardized tests, this booklet presents a practical, understandable rationale for the development and use of both kinds of testing in the educational setting. Section I discusses the functions of measurement, offering a rationale for assessing the potentialities and achievements of students. Section II addresses the question of the purpose of measurement, arguing that the major decision regarding a college's testing program is determining what to measure. Section III provides instructors, counselors, and administrators with a review of some of the generally accepted assessment devices for aptitude and achievement that are in general use in the schools. Section IV focuses on the development of the classroom test, including information on test organization and individual differences in organization. Section V covers practical considerations in testing, such as the consistency of testing and the appropriate time for testing. Section VI discusses problems in measurement, while section VII considers the selection of test items and types of items (i.e., multiple choice, matching, true-false, and essay). Section VIII discusses the principles for developing test items offering examples of good and poor items of each type. Section IX covers organizational components of classroom tests, including format, arrangement and ordering, correct response distribution, scoring, directions, correcting for guessing, and pretesting. Selected sample tests and a glossary are attached. (JMC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED332751

College of New Caledonia. Centre for Improved Teaching.
Prince George, British Columbia

APPROPRIATE

TESTING

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. Plett

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Duane O. Rubadeau

William A. Garrett

Ronald J. Rubadeau

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

College of New Caledonia Press
1990

JC 910270

BEST COPY AVAILABLE

APPROPRIATE TESTING

Duane O. Rubadeau
Co-ordinator, Centre for Improved Teaching
College of New Caledonia
Prince George, British Columbia

William A. Garrett
Director, Centre for Improved Teaching
College of New Caledonia
Prince George, British Columbia

and

Ronald J. Rubadeau
Director of Special Services
School District #23
Kelowna, British Columbia

© 1982, 1990, College of New Caledonia Press

Copyright 1982 by Duane O. Rubadeau, William A. Garrett and Ronald J. Rubadeau, Prince George, British Columbia, V2M 2S8. Reprinted 1990 by College of New Caledonia. All rights reserved.

ISBN 0-921087-16-0

Additional copies available from:

College of New Caledonia Press
3330 - 22nd Ave.
Prince George, B.C.
V2N 1P8
Phone (604) 562-2131

TABLE OF CONTENTS

Section		Page
	Preface	i
I.	The Function of Measurement	1
II.	The Purpose of Measurement	3
III.	Selection of Standardized Measurement Tools	6
	General Aptitude Tests	6
	General Achievement Tests	8
IV.	Developing the Classroom Test	10
	The Test as a Sample of Behaviour	10
	Test Organization	11
	Individual Differences in Organization	11
V.	Practical Considerations in Testing	13
	The Consistency of Testing	13
	When Do We Do the Testing?	14
VI.	Problems in Measurement	16
VII.	Selection of Items - Types of Items	19
	Multiple Choice Items	20
	Matching Items	21
	True-False Items	22
	Completion Items	24
	Essay Items	24
VIII.	Principles for Developing Test Items	26
	Multiple Choice Items	26
	Matching Items	43
	True-False Items	50
	Completion Items	60
	Essay Items	65
IX.	The Organizational Components of Classroom Tests	69
	Format of the Items	69
	Arrangement and Grouping of Items	70
	Correct Response Distribution	71
	Scoring Procedures	72
	Test Directions	73
	Should We Correct for Guessing?	74
	Comments about Pretesting	74
X.	Bibliography	76
XI.	Appendices	79
	Appendix A - Test Publishers Directory	80
	Appendix B - Selected Tests of General Aptitude	82
	Appendix C - Selected Tests of General Achievement	84
	Appendix D - Selected Tests of Reading Achievement	85
XII.	Glossary	87

PREFACE

The purpose of this book is to present a practical, understandable rationale for the development and use of both standardized and instructor-produced tests in the educational setting. We were going to use the term "school setting" but we found from experience that many of the tests employed in higher education show the lack of understanding of basic test and measurement principles as well.

Over the years we have seen a tremendous emphasis placed on the utilization of tests to the point, in many cases, where passing the test has meant more than the learnings that were supposed to be the objectives of the educational program. As time passed, the pendulum of test utilization in the educational setting swung in the opposite direction, with the hue and cry of why use tests at all. Both positions appear unreasonable, especially in the light of our recent interest in individualizing instruction and the need to provide information to the students regarding their capabilities and limitations. Even a casual conversation with students from junior high through college and university, reveals that many of them do not have a clue as to their assets and liabilities. As a result, we feel that a knowledge and understanding of the test and measurement principles involved in both standardized and instructor-produced tests would be of value to instructors at any level in the educational setting.

Each of these evaluation approaches have positive contributions to make to the educational setting. The instructor-produced test affords a flexible evaluation approach geared to the immediate instruction-learning situation. The standardized test, on the other hand, provides a long-range approach to achievement based on particular requirements of a course of study. This in turn, allows for comparison for a student's marks with the marks of students making up the standardized group.

In all probability, our orientation toward going overboard in the use of tests and the resultant backlash against the employment of tests, is that the standardized achievement test and instructor-produced test, collectively or singly, do not provide a complete picture for understanding the student. For example, they are not designed to measure potential or capacity. As a result, the students who have the highest scores on these tests do not necessarily have the greatest potential for the learning situation. Further, even within the area of achievement testing, the instructor-produced tests generally do not provide standards for comparing one student's performance with the performance of the average students within a particular class. In addition, we feel that many instructor-produced, as well as standardized tests, are of limited value in measuring student progress and especially, are of limited value in identification of strengths and weaknesses in the student's educational development.

Our orientation in this book is that in order for the educational setting to provide meaningful standards for comparison, sound bases for evaluating individual development and to evolve a diagnostic pattern of individual student strengths and weaknesses, instructors are going to have to make careful use of both standardized and instructor-produced tests.

As you might expect, there are several problems in developing this type of a testing program. First, is that most instructors tend to be whelmed, not overwhelmed, by the fantastic number of standardized tests in terms of their purposes and functions. Another problem is that available material about tests and testing is usually rather meagre, or is located in a wide variety of periodicals and books. This may not be a problem if you are teaching in a metropolitan centre or near a college or university; however, if you happen to teach in metropolitan Horsefly or South Porcupine, you will probably understand why this could be a problem. For example, the most widely used sources of test information, the Buros Mental Measures Yearbooks, provide an extensive coverage of the numerous standardized tests, but are generally not readily available to the classroom instructor. Finally, many, if not most instructors really do not have the foggiest idea of how to construct a valid and reliable test

for class use. They use a pattern, such as the types of test items they were given as students, or they assume that the items provided by the textbook publisher must be the example of good items. The first instance serves to compound the error, while the second case indicates our tendency to believe that because something is printed formally, it must be good.

As a result, there appears to be a need for an understandable book that can be used as a reference by instructors, administrators, and counsellors when faced with decisions regarding the application of test and measurement principles in the classroom.

It is the intention of the authors that this book be used as a guide in the development of a sound evaluation program utilizing both instructor-produced and standardized tests. As such, it suggests basic factors to be taken into account, rather than proposing "the way" to set up a testing program.

This book is divided into 12 sections: Section I deals with the functions of measurement; Section II includes material on the purpose of measurement; selection of standardized measurement tools including general aptitude and general achievement tests makes up Section III; the development of the classroom test including test organization and individual differences in organization comprise Section IV; Section V covers the practical considerations in testing; Section VI deals with the problems in measurement; selection of items and types of items are included in Section VII; Section VIII is a discussion of the principles for developing test items; the organizational components of classroom tests are covered in Section IX; Section X contains the bibliography; Section XI contains the appendices; and Section XII is the glossary.

We would like to acknowledge our indebtedness to the students and instructors who gave generously of their time and effort to comment and criticize the rough drafts of the manuscript to arrive at the present form, especially we thank Mr. George Worobey and Mr. Michael Sharlow of School District #28, Quesnel, British Columbia. We would be very remiss if we did

not acknowledge the excellent job of typing, proofreading and general trouble-shooting by our friends and assistants Faye Polyk and Susan Jardine.

Duane O. Rubadeau

William A. Garrett

Ronald J. Rubadeau

September 1982.

SECTION I

THE FUNCTION OF MEASUREMENT

The function of measurement in schools today has to be geared to our ever changing philosophy of education. For example, modern education emphasizes the view that each student have the opportunity to progress at his own pace. If we consider education to be a continuous process, then the student's development should be guided with skillful instruction toward his maximum. To achieve this goal, we need to utilize periodic assessments of the student's abilities and academic progress.

In order to plan for and attain the educational objectives of the school, we need to know the potentialities and achievements of the students. For example, we need to know potentialities and achievements to answer the following questions: What developmental progress is the student making under the current educational program? What kind of curriculum should the school have for the student? Which instructional methods appear to be most effective? Are there special strengths and weaknesses that will determine the organization of the student's educational program? Hopefully, a well-organized, understandable program will provide results enabling the school staff to begin to answer these questions.

Standardized and Instructor-Produced tests are utilized by those involved in the educational sequence. The results of these programs should provide feedback to the student, in terms of greater insight into his needs, achievements and potentials. In addition, much useful information can be gathered for the counsellor and instructor for helping the student to adjust and progress toward goals in line with his capabilities. Further, results of a good

testing program may be utilized in terms of developing broad educational program planning.

An important point we need to take into account is that test results must be supplemented with information about the student's background, home, health state, adjustment and so on. Providing experience and good judgment are involved in analyzing test results, we can obtain some fine insights into understanding each student as an individual. Our contention is that it is only when information about each student as an individual is available, that an adequate educational experience can be developed for each student.

SECTION II

THE PURPOSE OF MEASUREMENT

The first of a series of problems facing school systems when dealing with a testing program, is what to measure. Making up and administering instructor-produced tests is not, unfortunately, considered to be much of a problem; however, in dealing with standardized tests we occasionally find a school system utilizing standardized tests - because they have always used them. The point, a test should not be utilized, no matter how good it is, unless it serves some specific purpose. Equally, a testing program is useless if the results are filed away, or if it is installed without the cooperation and understanding of the instructors, or if its purpose is eyewash for the parents. For a testing program, or any other kind of program for that matter, to be a useful part of the educational process, all members of the educational hierarchy, instructors, counsellors and administrators should plan and develop the program together, hopefully basing the testing program on the educational objectives of the school system.

The major decision regarding the testing program then, is determining what to measure. This aspect of program development is often fun and games, as it calls for a review of the objectives of the school. After the objectives are reviewed, it is necessary to translate them into specific behavioural terms that can be measured objectively. The purpose is to move from the objectives, which often tend to be sweeping generalities, to the concrete, measurable behavioural statements.

After the educational goals have been defined, the remaining job is to determine what information is needed for each student, to evaluate and plan his or her school program. As a result, the testing or measurement program that is

developed will probably be unique for each particular school system. What has happened all too frequently in the past, however, is for the school system to adopt a textbook recommended program or even worse, to copy a program in use at another school. This again points up the need for utilizing the basic principles in the development of a sound measurement program. The general idea is that a measurement program that is highly effective and efficient in one school system may be a complete disaster for another system.

Whether or not the evaluation of the objectives and their translation into objective measurable behavioural statements has been completed, the most common procedure for a minimal testing program is to include two basic types of measures: measures of general aptitude (measures of intelligence), and measures of achievement in specific educational areas. With the general aptitude measures, we will obtain an indication of the students' potential. The measures of achievement on the other hand, provide an indication of how well each student has benefited from the instructional-learning situation as well as how capable he will be of learning. The combination of achievement and aptitude results should indicate to us the general type of instructional-learning situations that will be of greatest benefit to the student.

When a school system has special problems, and these appear more frequently as time passes, or when funds are available, a more elaborate measurement program may be developed. For example, the school system may choose to adopt various types of instruments such as interest inventories, personality and/or adjustment inventories (hopefully with parental consent), and special aptitude tests to aid in providing additional information for the individual student. The interest inventories assess the student's preference for various vocational and academic types of activity. The personality and/or adjustment inventories indicate typical response patterns as well as possible behaviour problem areas. The tests of special aptitude evaluate the student's learning potentials in special areas, such as mechanical ability, music, language, or mathematical reasoning.

The extensiveness of the measurement program for any school system, then, will depend upon the variety of objectives to be attained. However, when we come down to the nitty-gritty, it will be the pragmatic factors such as the budget, time, and trained personnel available that will determine the number of areas which will be evaluated. As a result of these practical factors, most school systems have to be satisfied to assess in the areas of achievement and general aptitude. Further, the utilization and interpretation of the personality measures necessitates the need for understanding the apparently nebulous results, as well as a large scale program to educate the parents as to the value of such instruments. These factors alone tend to negate their inclusion as part of a measurement program. This does not preclude the addition of these specialized assessment devices to a school program when trained personnel are available to insure proper use. However, for our purposes, we will limit our discussion to the areas of achievement and general aptitude as they are the foundation of the instructional measurement program.

SECTION III

SELECTION OF STANDARDIZED MEASUREMENT TOOLS

The purpose of this section is to provide instructors, counsellors, and administrative people with a listing of some of the generally accepted assessment devices that are in general use in the schools. The need here is to determine which of the tremendous number of available tests a school system should select in order to achieve the purposes of their measurement program. Most school personnel have little or no problem in determining the areas where measurement is needed. The difficulty is in deciding which specific measuring instruments to use.

With the large number of standardized tests available on the market, it would be a task way beyond the scope of this book to attempt a comprehensive listing of these tests. Instead, we have selected several tests for each area that have been utilized successfully in school measurement programs. This selection is not to be construed as meaning the tests omitted from the list are not good. Appendix A lists the addresses of the major test publishers. For critical reviews as well as additional information about tests, the Buros Mental Measures Yearbooks are excellent. Also see Appendix E.

General Aptitude

Before we get involved with specific types of general aptitude or intelligence tests, we have to decide on the type of measure we want to attain. That is, general aptitude measures that yield a single score as an overall measure of ability or the general aptitude measures that yield a total plus subtest scores.

The general aptitude tests where a single score is provided, is usually administered in one of two ways. It may have a series of separately timed

subtests or have a single time limit. Examples of tests with a single time limit are the Otis-Lennon Mental Ability Test and the Henman-Nelson Tests of Mental Ability. Examples of tests with separately timed subtests are the Differential Aptitude Tests and the California Test of Mental Maturity.

In a number of instances, tests yielding a single score do not provide enough of a description of particular aspects of the students ability. As a result, school personnel may find tests yielding several subscores or a profile to be more useful for determining guidance and instructional approaches. It should be pointed out at this time, that the items on both types of tests are very similar. The difference between them is a matter of the organization or grouping of items in order to provide subscore values as well as a total score. When subscores are available, school personnel may be able to utilize them to indicate specific types of student abilities to plan more adequate educational programs. Examples of tests that provide separate norms for subtests are the Cooperative School and College Ability Tests with verbal and quantitative scores and the Academic Promise Tests with language and nonlanguage scores.

Through the use of factor analytic techniques the multifactor tests were developed. The purpose of multifactor tests is to affect a separation of general aptitude into its basic components. These basic components, called the Primary Mental Abilities, are relatively independent, unitary factors such as verbal reasoning, spatial visualization and number facility. Two examples of multifactor tests are the SRA Primary Mental Abilities and the Flanagan Aptitude Classification Tests. When the school employs a multifactor test, information about the student's specific assets and handicaps becomes readily available, thus aiding the guidance and instructional planning program.

In summary, the advantages of using the single score tests in school measurement programs is that they provide an overall estimate of student general aptitude. Further, the single score tests are less expensive in both time and money. The major advantage of the multifactor and two-score tests for school measurement programs is that they provide a more extensive

understanding of student abilities and potentials in specific areas. A brief listing of various types of general aptitude tests will be found in Appendix B.

Achievement

Achievement tests are usually of two common forms and two general types. The two forms are: the single booklet consisting of a battery of tests, with each test covering a different aspect of achievement; and the separate tests for each area of achievement. The two types of achievement tests are those measuring content and those measuring application. The measures of content are the traditional variety of achievement tests. In effect, they measure student performance on specific content from the various subject matter fields. Subject areas such as arithmetic, science, social studies and reading are common components of the traditional achievement test batteries. Examples of the content type tests are the California Achievement Tests, Metropolitan Achievement Tests, and SRA Achievement Series.

The second type of achievement tests, the applied variety, are oriented toward the measurement of understanding of material in a broad area, compared to specific content in a given subject area. One of the most important goals in the educational process is to teach students to apply the material they have learned. The overall emphasis in this approach is general educational development. The areas that are typically measured on the applied type of test include interpretation of material in the biological and physical sciences, skill in quantitative thinking, data analysis, problem-solving, and understanding social concepts. Examples of the applied type achievement test are the Iowa Tests of Educational Development and the Canadian Tests of Basic Skills.

One of the difficulties commonly encountered by school personnel in the selection of an achievement test is that they may not be aware of the existence of the applied type of test. This is not to be taken as a snide remark, but rather, is intended to point up the similarity of the two types of achievement tests. For example, with a very thorough item inspection of the achievement tests, many similarities will be observed. Further, the popular approach of screening a list

of test titles may not be a very fruitful approach to determining the degree to which the tests measure different aspects of achievement.

The educational philosophy and the direction and organization of the measurement program of a school system are the obvious factors determining the choice of specific achievement tests. It is therefore very important to have a close similarity between test content and curriculum in the choice of a specific achievement test. For example, if the general educational development of the student is a major objective of the school program, assessment of specific content areas would not be as important as the measurement of ability to apply what had been learned. A brief list of commonly used achievement tests is included in Appendix C.

The ever-increasing pressure from various groups to provide more-and-more individualized instruction for students has generated the need to develop tests to obtain information regarding achievement in specific areas. Probably the most common of the lot are tests of reading ability. As most educational systems consider reading to be the basic educational skill, they want more information about the student's reading ability than that provided by the scores on the reading area of general achievement tests. An abbreviated list of reading tests is therefore provided in Appendix D.

SECTION IV

DEVELOPING THE CLASSROOM TEST

The Test as a Sample of Behaviour

Test Organization

The Test as a Sample of Behaviour

A test is nothing more than a sample of behaviour. As such, the items used to test a particular aspect of behaviour should be selected from all possible items that might be asked about that particular aspect of behaviour. Student performance on the sample of test items provides the basis for the instructor's generalization regarding the progress the student has made in the total area from which the sample is selected. Probably the most important element necessary for the instructor's generalization to be valid, is that the sample of test items is a fair and representative sample of all possible test items, related to a particular area.

Prior to the development of test items, the instructor should have in mind a clear and concise idea of the purpose of the test. That is, what content or skills are to be measured and how much of the test should be devoted to each of these areas? Perhaps this is an obvious point, however, all too often the test is comprised of a poor sample of items, which may provide a misleading picture of the student's progress in a particular area. As a result the instructor should have a plan or design for test organization in mind.

Test Organization

Test organization can be accomplished most readily if the instructor develops an outline of what the test is to cover. The outline should contain the various objectives to be measured and in the order of their importance, a list of the areas the test is to cover. Further, if the test is to measure the behaviours it is intended to measure, the instructor has to determine how much weight to assign to each of the objectives and areas of the test.

The need for such care in test development is to assure a fair and representative sampling of items so the test score will be a valid indicator of student performance. The general idea is that when the test is developed in accordance with an organized outline, the probability of the test score being a useful indicator of student behaviour will be enhanced greatly.

As an example of this type of approach we might take the development of a test in Social Studies. The outline we develop must first list the objectives to be measured. In the Social Studies area, our objectives might be the economic, political and social aspects involved in the western migration of the peoples in Canada. In addition, we might also list the topics to be covered, the chronological order of events and the skills to be assessed. Next, we have to determine the importance of each of these factors in order to assign relative weightings to each of these factors.

Individual Differences in Organization

Perhaps due to the lack of flexibility in many educational programs to which instructors have been exposed, there is a tendency on the part of many instructors to look for the organizational method to follow in test construction. The point, there is no single test organization pattern that is the best for all instructor-produced test construction. The idea is that in each subject matter area the test organization pattern will depend on the objectives to be evaluated. Further, each of the instructors, even those in the same area, probably will have a different emphasis as they deal with the subject matter. The differences in

emphasis will lead to greater flexibility in the test organization patterns. When the same instructor is dealing with student groups of different abilities or interests, still other test organization patterns would have to be used. Finally, the scope of the test will affect the organizational pattern. For example, making up the final examination will call for a different organization than making up a twenty-minute or an hourly test.

The main idea related to test organization is that if the test is to be worth the time spent to develop and administer it, the test should be organized so it will provide the information for which it was intended. Thus, every test should be developed from a carefully organized test pattern.

SECTION V

PRACTICAL CONSIDERATIONS IN TESTING:

The Consistency of Testing

When Do We Do The Testing?

The Consistency of Testing

Probably the most valuable aspect of testing is derived from a regular testing program. That is, instructional and administrative planning have the greatest possibility of being effective when they are centered around consistent measurement of the development and growth of the student. For example, the student's knowledge and understanding of his capabilities is likely to be enhanced when based upon an indication of progress made during a particular time period. If the time periods are of equal duration, the student would have a reasonable indication of long-range progress.

The general idea is that a testing program that is aperiodic may provide solutions for the more immediate problems the student encounters but this type of approach to testing is usually less effective in terms of sorting out the persistent long range problems. Further, the aperiodic testing program would be of limited value in a school system where the educational philosophy is oriented toward individualization. By using different levels of the same test throughout several grades, a sound testing program can be developed. In addition, the utilization of aptitude and achievement tests that have been standardized on the same students, will yield more useful results.

When Do We Do The Testing?

This will, of course, depend on whether we are evaluating aptitude or achievement. The aptitude or intelligence test, for example, would probably yield the most useful results if it was administered very early in elementary grades, then again in the intermediate grades, and finally in high school, with no more than a three-year period between testing. From such an aptitude testing program, the estimates of intelligence would be valuable for planning instructional programs for students. Also, a group approach to aptitude testing would be valuable for screening out those students needing additional assessment by the school psychologists or psychometrists.

While the establishment of an aptitude testing program appears to be simple and straight-forward, several precautions are especially important when utilizing the tests at the elementary school level. For example, developing rapport, and insuring the student is motivated and doing his best on the test, are factors which have a tremendous effect on the student's performance. Another important factor is that a single score from a group aptitude test should not be the sole determiner of the student's educational program. There are two main reasons for not utilizing the single aptitude test score in prescribing educational direction. First, group aptitude test scores are of limited value when dealing with any individual's score. Second, often there are striking changes in an individual's scores over time. These changes are attributable to such factors as emotional problems, health state, or reading problems.

The situation for establishing an achievement testing program should be based on a yearly testing procedure. The question of when the testing should be done during the academic year will depend on what you are intending to do with the results. One of the newer trends to develop has been the tendency to test in the Fall, and thus utilize the test results for instructional value. That is, testing at the beginning of the school year allows for instructional as well as administrative decisions regarding direction, placement and progress for the student - allowing for further individualization in planning.

If the achievement testing is centred around a Spring testing, the value of the program will be for its aid in Fall placement. Whenever the testing program is put into operation, it should be done on a yearly basis and the results should be available for decision-making at that time, rather than a delay of several months. This, of course, calls for some serious thought in order to organize and accomplish the goals of the program.

SECTION VI

PROBLEMS IN MEASUREMENTS

Most of the commonly used psychological tests have adequate reliability and validity; however, they are still a long way from being perfect. As a result, the interpretation of results and the uses to which the tests are subjected can legitimately be questioned. It is only reasonable then, that we view a student's test score as an estimate, rather than as the final and unchanging indicator of ability.

Some of the problems encountered in psychological measurement that must be taken into account are: (1) Test scores, in individual cases, may be based on false or erroneous assumptions. (2) Temporary or accidental variations in the individual may have a tremendous effect on the test scores. And (3) Emotional trauma, inadequate motivation or poor physical conditions can lead to spuriously low scores. In addition, in achievement testing, if a school curriculum differs greatly from the curriculum on which the test is constructed, the test will not yield a fair estimate of student achievement. Further, if the student has a handicap in dealing with verbal material, such as English not being the first language, utilization of a test of verbal aptitude will probably not provide a fair estimate of verbal ability. While these problems do affect test results, it is often what we do with the results that produces problems in the area of measurement.

One of the best examples of the treatment of test results that produces problems occurs when the school adopts grade or age norms as achievement standards. By definition, a norm is the average performance on a test by a group of students in a particular grade or of a particular age. Further, if a large number of students make up the norm group, half the students will be above the

average mark and half will be below the average mark. Thus, when the school sets the norm as the average performance, approximately half of the students in any normal classroom will not be able to reach the norm figure.

Instead of utilizing norm group data, the standard for achievement for any student should be based on his own capacity for achievement. Hence, a student in the fourth grade having third grade mental ability, would be expected to achieve third rather than fourth grade performance in reading.

Another problem in dealing with the norm as a behavioural standard is actually the converse of the problem previously discussed. That is, when a student achieves a high standing in terms of the norm, it does not necessarily mean that the student has achieved the educational objectives of the school. For example, if the students in a particular school received a straight rote memory course in Social Studies, they might well score above the norm on an achievement test in Social Studies, yet might have done a very poor job on the areas of the test that deal with understanding of meanings and relationships.

In summary, what we have attempted to portray in this section of the text is that psychological and educational testing is only the means to an end. In no way is it the end in itself. When tests are used properly, the results can be utilized to give direction and guidance to students. As a result, a successful performance or a failing performance on a test should not be allowed to become the all-important factor.

The primary purpose of any psycho-educational testing program should be for information gathering, to be used to further individualize the students instructional program. A testing program should not be the primary instrument in instructor evaluation or the instrument to determine promotional status. It is not uncommon, however, that instructors and students wrongly assume that a high test score is the primary educational objective.

With any psycho-educational measuring device, the test should be keyed to the instructional objectives or the curriculum. Tests, whether standardized or

instructor-produced, are only a part of the educational process, and are useful only insofar as they help to serve the student.

SECTION VII

SELECTION OF ITEMS - TYPES OF ITEMS

Developing Test Items

Organizational Components of Classroom Tests

Procedure for Reviewing Classroom Tests

For many years the debate has been raging among academicians as to the relative merits of the various types of test items. After all the dust had settled, (the argument is not over, it is just that most educators and psychologists have become preoccupied with other endeavours) the general idea is that there is no general answer to the question of which type of item has most merit. For example, are true-false items better than multiple choice items? Or are objective-type items more reliable than essay-type items?

Each of the items that will be examined in this section will have its own advantages and strong points, as well as its disadvantages and weak points. As you deal with the various types of items, the major concern you should keep in mind is how well the item is constructed, not the type.

Types of Items

Test questions are broadly classed as either objective or essay. Although it is occasionally difficult to differentiate between these broad classifications, the objective-type item tends to limit the student's response to a word, phrase or symbol which would be designated as the correct response. The two main subdivisions of the objective-type items refer to selecting the correct response versus supplying the correct response. The selection-type of

item includes multiple-choice, matching and true-false. The supply-type is best exemplified by the completion item. One of the major factors in dealing with the objective-type item is that they can be scored using keys, by people who are not necessarily specially trained for the task.

The essay-type item is usually reserved for the situation where the instructor is interested in the student's ability to express himself and organize material into a logical sequence. In the essay-type item, the instructor, as the expert in the field, must make a subjective judgement about the relative quality of the student's response.

At the present time there are many varieties of item styles available. When all of the variations of the standard-type test items are taken into account, the instructor has a great deal of freedom in choosing an item style that will allow for optimal measurement of the objectives for any type of course. The rest of this section will be concerned with a discussion of the relative merits of the more commonly used objective-type items such as multiple-choice, matching, true-false and completion items, and the essay-type item.

Multiple-Choice Items

The multiple-choice item consists of a stem and two or more responses. The stem is ordinarily an introductory statement or question. The responses are possible answers for the question or statement. For example:

The capital of Canada is:
a. Vancouver
b. Toronto
c. Ottawa
d. Montreal

The main advantages of the multiple-choice item are that they can be relatively free of ambiguity and they are applicable to a wide variety of subject areas. The major factor for avoiding ambiguity, is that the multiple-choice item involves the selection of the best answer available. Thus, the multiple-choice item avoids the ambiguity that is encountered when dealing with absolutes. The

wide applicability of the multiple-choice format allows for testing everything from knowledge of specific details to application of what has been learned to reasoning ability and understanding of fundamental relationships.

One objection to the multiple-choice format is often voiced by students and occasionally by instructors. This objection is centred around the guessing factor in multiple-choice items. While a number of students refer to these items as multiple-guess, the guessing factor is reduced greatly on a good multiple-choice test. For example, on a 25-item multiple-choice test, with five responses for each item, the probability of getting a mark of 70 percent could occur by chance alone is about one time in one million. Thus, in order to do well on a multiple-choice test, knowledge rather than guessing is the major factor in performance.

Matching Items

The matching-type item format is composed of two lists. The first is a list of the premises, the second is a list of possible answers. The student's task is to match one of the responses to each premise. For example:

> In the space to the left of each investigator in column A, write the letter from Column B that is in front of the contribution they have made.
>
> | | | Column A | Column B |
> |---|---|---|---|
> | _____ | 1. | Pavlov | a. Conditioned neuroses |
> | _____ | 2. | Thorndike | b. Operant conditioning |
> | _____ | 3. | Liddell | c. Reinforcement |
> | _____ | 4. | Skinner | d. Condition reflex |
> | _____ | 5. | Kohler | e. Insight learning |

The advantages of the matching-type item are: ease of scoring, objectivity in scoring, guessing is kept to a minimum, reading time is small, allowing for a relatively large number of items in a limited time period. The main disadvantages of the matching item is that much of the flexibility found in the other objective items is limited. That is, this type of item is useful when the material used can be listed but would not be overly valuable when the topics do not lend themselves to the listing procedure. In the previous example provided,

all of the premises and responses are from the area of learning in the field of psychology. They would be considered to be homogeneous in nature, i.e. any one of the responses could be a plausible answer for any one of the premises. Where the premises are heterogeneous, the possible interchange of responses is limited, allowing for guessing of the correct response. For example:

> In the space to the left of each of the persons listed in Column A, write the letter from Column B that describes this person's work.
>
> | Column A | Column B |
> |---|---|
> | 1. Hemingway | a. musician |
> | 2. Trevino | b. singer |
> | 3. Mozart | c. author |
> | 4. Nuryev | d. dancer |
> | 5. Lightfoot | e. movie star |
> | | f. golfer |
> | | g. entertainer |

True-False Items

The true-false item is nothing more than a statement that the student must judge as true or false. A variation of the true-false item is where questions are substituted for the statements and the student responds yes or no to the question. For example:

True	False	The largest province in Canada is Ontario.
Yes	No	Is British Columbia known as the dairy province?

The advantages of the true-false items are: objectivity in scoring, ease of adaptation to the classroom setting, and the broad sampling of course content, as many items can be completed in a relatively short period of time.

The disadvantage of the true-false format is: the tendency to use these items to assess for memory of specific factual material. The major reason for this state of affairs is that the true-false item must be based on material that is

absolutely true or absolutely false. In most academic areas there would be relatively few statements that would fall into this category. For example, situations that call for the student to evaluate, explain or develop generalizations, cannot be handled via the true-false format. In addition, a real drawback to the true-false item is the care and time that must be taken so that the student responds only to the item as stated, rather than attempting to bring in other data such as qualifying statements or single instances from their own past in order to answer the item. A further limitation is the guessing factor in the true-false format. To reduce the chance factor in getting a high mark on a true-false test, more items must be added to the test. However, as more items are added the fatigue factor can serve to reduce the effectiveness of the test. As an example, if we have a true-false test of 50 items, the probability of the student getting a mark of 70 percent by chance alone is 1:350; by doubling the number of items to 100, the probability of a 70 percent mark will be reduced to approximately 1:10,000.

Before going on to the last of the objective items, we might examine a modification of the true-false item that has become popular with a number of instructors. For example:

> Directions: If the statement is true, circle the T, if the statement is false circle the F and write the word that should be substituted for the underlined word that will make the statement true.
>
> T F The Thames River runs through <u>Ireland</u>.

The modified true-false item does allow for a bit more flexibility in determining whether the student knows the correct response when the answer is false. While guessing is somewhat reduced in this approach it is not eliminated.

Completion Items

In effect, the completion-type item is a combination of the objectivity found with the multiple-choice item and the recall from the essay item. In the completion-type item, the student responds by supplying the portion of a statement that is missing. For example:

> The highest mountain range on earth
> is the _____.

The advantages of the completion items are: the virtual elimination of guessing, as recall rather than recognition is measured, scoring can be done in a relatively objective and rapid manner, and sampling of a broad area of content can be accomplished quite easily.

The major disadvantage of the completion-type item is that like the true-false item, it is generally limited to the measurement of facts and general information. The reason for this restriction is that the completion item must be couched in a rather careful style to eliminate a wide variety of responses. Hence, construction of completion items becomes restricted to the measurement of factual material in everyday practice.

Essay Questions

The value of the essay-type item is that it is free of most restrictions which must be taken into account with many of the objective-type items. The essay item is based completely on recall of material from the student's experience, with no cues available. The primary purpose of the essay item is that it allows the instructor to assess the student's ability to organize, analyze, evaluate and develop solutions to problems. In so doing, the essay-type item appears to evaluate some of the higher level cognitive abilities of the student.

Probably the major disadvantage of the essay-type item is the difficulty in establishing reliability of test scores. Several error sources contribute to the unreliability of essay test scores. One problem is the limitation in terms of

content sampling. The difficulty here is that usually the essay test is limited to very few questions, which tends to restrict the sampling of course content. For example, a student has a good background in the historical development of a social condition but a relatively poor understanding of the political and social implications related to this condition. If the essay test is oriented toward the historical aspect, this student has it made. However, if the test deals with the political and social factors, he has had it.

The other common problem when dealing with essay test unreliability, is the subjectivity involved in scoring. Much of the negative reaction toward essay exams appears to be justified, as evidence has been put forth to show the variation in scoring by the same teacher at different times during the day, let alone the differences between teachers, in scoring essay exams. The problem of lack of reliability when scoring essay tests seems to be centred around the development of an adequate scoring key with definite assignment of points for the organization, analysis or problem-solving approach used by the student.

In spite of the problems inherent in scoring the essay-type examination, when the questions are adequately designed and a scoring procedure is developed which allows for some degree of objectivity, the essay item can provide very useful information for both student and instructor.

SECTION VIII

PRINCIPLES FOR DEVELOPING TEST ITEMS

The purpose of this section is to survey for the reader various aspects involved in the development of the several types of test items. The material included in this section is not to be construed as the last word in item construction, however, it will cover the most common problems in item construction, as well as to point out the major hazards to avoid.

The development of good test items is not an easy task. However, if attention is focused on the basic principles of item construction included in this section, much wasted time and effort can be eliminated and effective items can be written - items that will measure reliably and validly the various aspects of course content.

Although there is bound to be some degree of overlap in the principles in terms of specific item types, the approach used in this section will be to list the principles for each type of item, then follow through giving examples for each of the principles. The items will be covered in the following order: multiple-choice, matching, true-false, completion and essay.

Multiple-Choice Items

To say the least, the principles of good multiple-choice construction are extensive in nature. The list of principles:

(1) The item should be written in clear language, with the simplest possible vocabulary.

(2) Each item should be based on one central problem.

(3) The stem should be either an incomplete statement or a question.

(4) The stem should include words that would have to be repeated in each of the choices.

(5) The stem should have only one correct answer.

(6) The stem should contain a clear and complete statement of the central problem.

(7) Avoid excess or irrelevant material in the stem.

(8) Avoid the use of negative statements in the stem.

(9) All responses should be plausible to students lacking the ability or information being tested.

(10) Where possible, the responses should be arranged in a logical order.

(11) Responses should be placed at the end of the stem.

(12) The stem should avoid unintentional clues.

(13) The stem and response should be grammatically consistent and parallel in form.

(14) The responses should be independent and mutually exclusive.

Now to cover each of these principles in depth, with appropriate examples:

1. The item should be written in clear language, with the simplest possible vocabulary.

The idea here, is that sentence structure should be as simple as possible. As with all other types of items, the meaning of the multiple-choice item should be clear. That is, highly technical or extremely difficult vocabulary should be avoided. Further, whenever possible, the important aspects of the statement should be positioned at the beginning of the stem. Examples of how this principle should be put into operation:

Poor Item:

In the Prairie Provinces of Canada, the season of the year which the records indicate as having the greatest probability of rainfall is:

a. early spring
b. summer
c. fall
d. midwinter
e. late winter

Good Item:

In the Prairie Provinces of Canada, rainfall is most likely to occur in:

a. early spring
b. summer
c. fall
d. midwinter
e. late winter

In this example, the meaning of the stem was improved immensely by reorganizing the statement into clear, concise terms. Often, due to the poor construction of the item, the instructor may be measuring the ability of the student to decipher the stem instead of the student's ability to solve the particular problem.

2. Each item should be based on one central problem.

The multiple-choice item is most effective when it focuses the student's attention on one central problem. Here, it is not unusual to find items that are nothing more than a group of unrelated true-false statements having the stem as a common beginning.

Poor Item:

A test should:
a. predict behaviour
b. have practical significance
c. detect individual differences
d. classify persons accurately

The item is, in effect, dealing with four different problems. As such, this series of true-false items forces the student to determine which of the responses is more true than the other ones. To say the least, this is a difficult situation. Exposing students to a series of items of this type will usually result in student outcries along the line of ambiguity and unfair testing, which can easily generalize into comments about "poor instruction."

This item can be improved quite easily, if we supply additional qualifying information in the stem. This additional information allows the student to relate each response to the specific problem stated in the stem, rather than going through the process of weighing the relative truth of each response against the other responses.

Good Item:

In order to aid psychologists and educators in making decisions, a test should:

a. predict some present or future behaviour
b. have practical significance
c. detect individual differences
d. classify persons accurately

Another rather common difficulty involved in item construction that students find confusing is also related to the central issue problem. The confusion results when the item deals with two or more separate problems. That is, this type of item often involves diagnosis and prognosis, or diagnosis and treatment, or prediction and control.

Poor Item:

In Phenylketonuria, the individual:

a. suffers brain damage within the first two years
b. should be placed on a phenylalanine-free diet
c. shows lack of pigmentation in skin and hair
d. shows retarded mental and physical development

Here we have, diagnosis in terms of symptoms, prognisis, and treatment all involved in the single item. Generally, this type of item leads to confusion, as the subject does not have a clue to the main problem. This type

of item can be improved measurably by focusing on any one of the three areas - diagnosis, prognosis or treatment, thus, developing three items out of the original one.

Good Items:

Which is a symptom of phenylketonuria?
- a. lack of pigmentation in skin and hair
- b. allergic reaction to high protein food
- c. abnormal body development
- d. positive results on Guthrie's Blood Screening Test

The prognosis for phenylketonuria is:
- a. brain damage within two years without treatment
- b. microcephaly
- c. abnormal body development
- d. retarded mental but not motor development

The treatment for phenylketonuria is:
- a. a phenylalanine-free diet
- b. psychosurgery
- c. psychotherapy
- d. remove vegetables from the diet

3. The stem should either be an incomplete statement or a question.

The idea here, is that you should use the item stem style that seems most appropriate for the specific item you are constructing. That is, for some content, the question-type stem may be most appropriate. In other cases, the incomplete statement-type stem allows for the most adequate handling of the content.

Examples:

Which has the smallest brain?
- a. cretins
- b. microcephalics
- c. macrocephalics
- d. hydrocephalics

A test can be defined as a:

a. set of questions covering a specific content area
b. situation requiring problem solving
c. fixed procedure for comparing the behaviour of two or more people
d. task requiring a person to do his very best

Generally, the beginner will experience less difficulty and produce better items when the question-type stem is used. The primary reason for this state of affairs is that the test constructor must come up with specific choices to answer the question, whereas in the incomplete statement-type stem, much concern and effort has to be directed toward construction of the proper qualifying terms used in the stem. Other than that, there is really no difference between the question and the incomplete statement items as far as measuring effectiveness is concerned.

4. The stem should include words that would have to be repeated in each of the choices.

Here, the idea is that all aspects that the choices have in common should be transferred into the stem. There are two major reasons for this. First, when the common aspects of the choices are removed, it is easier for the subject to focus on the problem, and second, by removing the common aspects from the choices, the choices are easier to read and understand.

Poor Item:

Psychology is defined as:

a. the scientific study of our inner conflicts
b. the scientific study of our observed behaviours
c. the scientific study of our sensations
d. the scientific study of our psyche

Good Item:

Psychology is defined as the scientific study of our:

a. inner conflicts
b. observed behaviour
c. sensations
d. psyche

5. The stem should have only one correct answer.

What often occurs in the development of items, is to bring in material that is confusing or ambiguous. Hence, the need for this principle. When the content in the choices is so ambiguous that instructors in the field cannot agree on the correct choice, the item is obviously not going to be effective for measuring student behaviour.

Poor Item:

Galton is best known for:
a. research
b. developing statistical methods
c. the law of gravity
d. psychophysical methodology

Good Item:

Galton is best known for:
a. research on individual differences
b. developing statistical methods
c. the law of gravity
d. psychophysical methodology

The item above illustrates the way several possibly correct or at least debatable items can be eliminated as choices and only one correct response remains.

6. The stem should contain a clear and complete statement of the central problem.

The idea here, is that the item stem should get across the main issue. To accomplish this, the stem should put the problem in a clear and concise manner.

Poor Item:

Behaviourism:
a. sought objectivity
b. denied introspection
c. was mentalistic
d. opposed Gestalt Psychology

The student, confronted with this item, probably would not have a clue as to what the problem was or what the instructor expected. To improve this item to bring about better understanding, additional clarifying material is provided.

Good Item:

The main theme of Behaviourism was to:
a. seek objectivity
b. deny introspection
c. develop mentalism
d. oppose Gestalt Psychology

7. Avoid excess or irrelevant material in the item stem.

There are several reasons why item content should be confined to material relevant to the problem. First, excess verbal material will penalize those students who are slow in comprehension. Second, the excess verbal material would penalize the students who have limited ability in speed reading. Third, and most important, the excess verbal material may produce ambiguity and confusion as to what constitutes the problem.

Poor Item:

Your instructor has discussed several techniques that are commonly used to collect data, such as the experimental method, statistical method, field observations and the case study. The basic limitation of field observations is that they are:
a. unstructured
b. unstandardized
c. unreliable
d. invalid

Good Item:

The basic limitation of field observations is that they are:
a. unstructured
b. unstandardized
c. unreliable
d. invalid

8. Avoid the use of negative statements in the stem.

The common result of the use of negative statements in the item stem is confusion. The greater the complexity of the item content, the more difficulty the student will have in comprehending the problem. The difficulty is further compounded when double negatives are utilized.

Poor Item:

In a pure power test, test takers would not differ in the:
a. number of items attempted
b. number of items not attempted
c. number of items answered correctly
d. number of items not answered correctly

In the above item, the combination of negative statements in the stem and choices results in mass confusion for the subject.

Good Item:

In a pure power test, test takers would be similar in the:
a. number of items attempted
b. number of items answered correctly
c. percent of items answered correctly
d. time taken to complete the test

9. All choices should be plausible to students lacking the ability or information being tested.

The idea here, is to construct the choices so students with limited knowledge or those who have not studied sufficiently will find the choices equally attractive. What occurs quite often as the instructor is constructing the

item, is that two or three plausible choices are listed, then two or three ridiculous choices are added to give the desired number of choices.

Poor Item:

The branch of psychology most directly concerned with the emotionally disturbed individual is:

a. animal
b. clinical
c. industrial
d. aesthetic

Good Item:

The branch of psychology most directly concerned with the emotionally disturbed individual is:

a. social
b. clinical
c. abnormal
d. personality

10. Where possible, the responses should be arranged in a logical order.

When an item contains several choices dealing with the same concept, it is advisable to place them in some logical order, say from positive on through to negative, rather than to have them tossed together on a random basis.

Poor Item:

The figures below represent the correlation between scores on two tests. In which case can scores on one test be predicted most accurately from scores on the other test?

a. r = -.83
b. r = +.75
c. r = 0.00
d. r = +.50

Good Item:

The figures below represent the correlation between scores on two tests. In which case can scores on one test be predicted most accurately from scores on the other test?

a. r = -.83
b. r = 0.00
c. r = +.50
d. r = +.75

When the choices are numbers, two factors should be kept in mind. First, either the choices should be indicated by letters rather than numbers, or the numbers making up the choices should be written out. The reason is that students will often become confused when the choices which are indicated by numbers are also numbers. Second, it is good practice to arrange the choices in either ascending or descending order.

Poor Item:

The median for the following set of data is:
3, 1, 2, 2, 4, 1, 4

1. 2.5
2. 3
3. 2
4. 1.5

Good Item:

The median for the following set of data is:
3, 1, 2, 2, 4, 1, 4

a.	3.0		a.	three
b.	2.5	or	b.	two and five-tenths
c.	2.0		c.	two
d.	1.5		d.	one and five-tenths

By using letters to designate the choices, the confusion between the two sets of numbers, as in the poor example shown, is eliminated.

11. Responses should be placed at the end of the stem.

One of the common problems in the construction of classroom tests is to have the student insert the correct choice into the middle of the stem statement. This is very confusing to students, as they cannot understand the problem without looking at the choices and when they look at the choices they do not have a clue as to what the problem might be.

When the choices are placed at the end of the stem statement, the student is better able to understand the problem and respond accordingly.

Poor Item:

A(n) _____ scale has an absolute zero point and equal units of measurement.

a. Nominal
b. Ratio
c. Ordinal
d. Interval

Good Item:

Which of the following scales has an absolute zero point and equal units of measurement?

a. Nominal
b. Ratio
c. Ordinal
d. Interval

12. The stem statement should avoid unintentional clues.

It is quite common on classroom tests, to find the instructor has inadvertently provided clues for the student that allows for solution of the problem, even though the student may not really comprehend or understand the concept being evaluated. The student's task is to associate each of the choices with the stem in order to arrive at a solution to the problem. When unintentional clues such as: repeating key words in the choice and stem, inconsistent grammar, or unusual length of the choice are provided, the item is not measuring the students' understanding of the concept.

Poor Item:

The first psychology laboratory was started in Germany by:
a. Galton
b. Wundt
c. Jones
d. Smith

Good Item:

The first psychology laboratory was started in Germany by:
a. Fechner
b. Wundt
c. Wertheimer
d. Kohler

In the poor example above, the student would have a good chance of answering the item correctly on the basis of selecting the only choice with a German name. This clue can be offset, as in the better example, by having all German names for the choices.

Another of the common ways in which clues are inadvertently supplied to the student is when the correct choice is much more detailed and longer than the alternatives.

Poor Item:

The distinction between an aptitude and an achievement test is based on:
a. the type of items
b. the purpose for which the test is administered
c. the time limits
d. item difficulty

Good Item:

The distinction between an aptitude test and an achievement test is based on:
a. the kind of items that make up the test
b. the purpose for which the test is given
c. the time limits that the subject is given
d. the difficulty level of the items

By eliminating the unusual length or shortness of the correct response, you improve the possibility of obtaining a correct response on the basis of knowledge or understanding.

13. The stem and responses should be grammatically consistent and parallel in form.

The emphasis here, is to develop a combination of stem and responses that go together to form a logical and clear pattern of thought. When this idea is kept in mind the student will be able to grasp the intended meaning of the item much more readily. When the test constructor neglects to keep this idea in mind, the student may have extreme difficulty, as the statements are not logical or expressed clearly.

Poor Item:

A selection decision is:

a. a counsellor advises a person regarding the vocational programme he seems best suited for
b. a student decides whether he should accept a job offer with a company
c. an institution accepts some applicants and rejects others
d. employees may be placed in particular jobs depending on results of aptitude tests

In this example, the lack of grammatical consistency is a source of confusion for the student.

Good Item:

A selection decision is one in which:

a. a counsellor advises a person regarding the vocational program for which he seems best suited
b. a student decides whether he should accept a job offer with a company
c. an institution accepts some applicants and rejects others
d. employees may be placed in particular jobs depending on results of aptitude tests

When the alternatives are made to be consistent with the stem, as in the above example, the item becomes much more meaningful.

14. The responses should be independent and mutually exclusive.

The main idea here, is that when the responses are interrelated in terms of meaning, they provide cues to help the student eliminate the wrong responses. The net result of interdependence of responses is to decrease the number of effective responses. This, in turn, leads to a reduction in reliability. For example, on the instructor-produced test, it is not unusual to encounter items where two, three or all of the responses available to the student cover the entire range of possible responses. When this occurs, only one of the choices will be acceptable to the student as being correct. Further, occasionally the student will encounter an item where one of the responses includes two or more of the other responses, which the student can eliminate by deduction, rather than by knowledge or understanding of the material.

Poor Item:

Learning, as differentiated from maturation, refers to:

a. the result of practice
b. the elimination of practice
c. neurological change
d. irreversibility of behaviour

In this example, the first two choices cover the range of possibilities, that is, either a behavioural change occurs or it does not occur. This item can be improved by further qualification in the stem and making the responses independent.

Good Item:

The part of the definition of learning which distinguishes learning from maturation is the part which says that learning is:

a. a change in behaviour
b. fairly permanent
c. the result of practice
d. irreversible
e. related to neurological change

The following example indicates another type of response interdependence:

Poor Item:

What percent of the scores in a normal distribution are lower than one standard deviation below the mean?

a. less than 8%
b. less than 16%
c. more than 32%
d. more than 50%

The student utilizing the process of elimination, can reduce this item to a two-choice item. That is, if A is correct, then B is also correct. Conversely, if D is correct, C is also a correct choice. Thus, the student that is wide awake and thinking is able to eliminate choices A and D as possible correct responses. Items of this variety are readily improved by making the choices independent. For example:

Good Item:

If test scores are distributed normally, what percent of the scores will exceed a score falling one standard deviation below the mean?

a. 16%
b. 34%
c. 68%
d. 84%

Before going on to deal with the other types of items, we should mention one other approach that may be encountered on multiple-choice type

tests - the universal "all-of-these" or "none-of-these" responses. In effect, they supposedly allow for some of the flexibility that is commonly found on the essay-type test. While some proponents of the "all-or-none" response feel they offer greater flexibility and therefore, allow for better evaluation of the students ability, in actuality these responses still appear to be measuring recognition ability rather than some other cognitive component. For example:

Poor Item:

A control group is used in order to:

a. hold all factors constant except the variable being studied
b. compare two different samples of the same population
c. increase the sampling to ensure greater stability in the results
d. none of these

What occurs most often in these items, is that the instructor had trouble thinking of another response, thus adding the "all-or-none" to stretch the possibilities. Before getting carried away and dismissing the "all-or-none" type of response as being utterly useless, we should point out that this type of response can be quite valuable in assessing the student's ability to recognize a situation where none of the alternatives can logically be selected. Tests of ability to analyze data provide a real fine example along this line.

Good Item:

An investigator finds a statistically significant correlation between variables X and Y. From this information, the most justified conclusion is that:

a. variables X and Y are strongly related to each other
b. variables X and Y are strongly related to each other, but in an inverse manner
c. variable X is a very good predictor of the values of variable Y
d. none of these

The generalization related to the "all-or-none" type of response is that they can be quite useful, depending on the type of material being evaluated.

However, they are commonly added on to the end of the list of alternatives as a means of providing more choices.

Matching Items

The principles related to the development of matching items will be listed first, then covered individually.

(1) Premise and choice lists should be limited in length.

(2) The directions for matching should be clear and concise.

(3) A set of matching items should be made up of related premises and related choices.

(4) Extra choices should be provided to reduce guessing.

(5) To avoid undue confusion, organize the lists of premises and choices.

Now to cover each of these principles in detail.

1. Premise and choice lists should be limited in length.

The general idea here, is to keep the lists of premises and choices to a length that can be easily handled by the student. The length of these lists will, of course, be dependent on the age and ability level of the students. It is not uncommon for the student to be faced with a list of 25 premises and 30 choices making up a matching set.

The major disadvantages of the long matching sets are: the great amount of time in the search for the correct choice; and the difficulty in developing a homogeneous set of premises and choices, that still allows for an adequate sampling of the course material.

A rule-of-thumb that may be applied when developing matching sets is that a set of 10 items would be the maximum. Preferably, the ideal set size would be from 6 to 8 items.

2. The directions for matching should be clear and concise.

Even though a matching set may be quite clear and straight-forward, it is a good idea to be as clear and concise as possible with the directions. For example, giving the student specific directions such as: literary works are to be matched with their authors, or political events are to be matched to the Prime Minister in whose administration they occurred, provides the student with an orientation that is very helpful to him. In the following sample item, the student faces a good possibility of becoming confused. Further, there are a number of students who will never develop a proper classification system by themselves.

Poor Item:

On the line at the left of each item in Column A, write the letter of the matching item in Column B.

	Column A		Column B
_____ 1.	John B. Watson	a.	Schooling dealing with phenomenology
_____ 2.	Wundt	b.	Developed psycho-analytic school
_____ 3.	John Stuart Mill	c.	Started school of behaviourism
_____ 4.	Freud	d.	Started school of Structural Psychology
_____ 5.	Wertheimer	e.	Associationist School

Good Item:

On the line to the left of each school of psychology listed in Column A write the letter of the psychologist from Column B associated with that school.

	Column A		Column B
_____ 1.	School dealing with phenomenology	a.	John B. Watson
_____ 2.	Developed psychoanalytic school	b.	Wundt
_____ 3.	Started school of behaviourism	c.	John Stuart Mill
_____ 4.	Started School of Structural Psychology	d.	Freud
_____ 5.	Associationist School	e.	Wertheimer
		f.	Adler
		g.	Horney

By making the basis for classification quite explicit in the directions, the item is improved quite nicely.

3. A set of matching items should be made up of related premises and related choices.

Another of the major problems associated with the development of matching items is to find premises and choices that are homogeneous, which will allow a meaningful basis for matching. When the items in each of the matching lists are heterogeneous, they can often be solved by students having only a vague concept, just on the basis of simple verbal associations. For example:

Poor Item:

	Column A	Column B
_____ 1.	Extraneous variable	a. Father of British Psychology
_____ 2.	Correlation coefficient	b. Started first psychology laboratory in Germany
_____ 3.	Sir Francis Galton	c. A method for determining the relationship between variables
_____ 4.	The Case History	d. Factors not studied directly, but could affect the results.
_____ 5.	Wilhelm Wundt	e. A method of collecting past history for a client

In this matching set, the items are so heterogeneous in Column A, that each item is obviously related to only one item in Column B. All of the other choices are very poor distractors.

A better example of a matching set is as follows:

Good Item:

	Column A	Column B
_____ 1.	Factors not studied directly, but could affect the results	a. Independent variable
_____ 2.	The events a science tries to predict	b. Extraneous variable
_____ 3.	The antecedent condition	c. Dependent variable
_____ 4.	A method for determining the relationship between variables	d. Experimental method
_____ 5.	The resultant response we are trying to measure	e. Stimulus

By utilizing choices that are homogeneous in nature, it is a rather simple matter to sort out those students who understand the concepts from those who do not understand the problem.

4. Extra choices should be provided to reduce guessing.

A good matching set has the advantage of reducing guessing, and hence providing a reliable measure of the students' understanding of the material. If we use a 12-item matching set and the student does not know the answer to one of the items, the chances of getting the right response by guessing is relatively small. A common difficulty encountered in instructor-produced matching sets, is to have the same number of premises and choices. If the student knows 10 of the 12 items, his chance of guessing the other two items is 50-50. As a result, it is a good idea to have two or three more choices than premises, hence, further reducing the possibility of coming up with a correct response by guessing. For example:

Good Item:

	Column A	Column B
_____ 1.	Factors not studied directly, but could affect the results	a. Independent variable
_____ 2.	The events a science tries to predict	b. Extraneous variable
_____ 3.	The antecedent condition	c. Dependent variable
_____ 4.	The consequent condition	d. Experimental method
_____ 5.	A method for determining the relationship between variables	e. Stimulus
_____ 6.	The resultant response we are trying to measure	f. Response
		g. Clinical method
		h. Observational method

5. To avoid undue confusion, organize the lists of premises and choices.

The common approach students use with matching sets is to read the first premise in the left-hand column, then go down the list of choices in the right-hand column until a plausible solution is found. As a result, each of the premises is read only once, while the choices may be read and re-read a number of times. To make the matching set as convenient as possible for the student, it is of value to arrange the lists in an organized form. That is, the longer statements should be utilized as premises and placed in the left-hand column. The choices should be short, concise statements such as names, dates, or places, and placed in the right-hand column. The reasoning involved in this operation, is that the short statements can be remembered and compared more readily than a set of complex, long statements. Further, the lists should be organized in a logical order, which further simplifies the matching procedure. For example, names can be arranged alphabetically and dates can be ordered chronologically.

The generalization that can be applied to the matching items is that the more simple the task for the student, the better the item will measure the behaviour being studied.

Poor Item:

	Column A	Column B
____ 1.	Sigmund Freud	a. The events from which the prediction is made
____ 2.	Correlation coefficient	b. The study of structures and functions underlying behaviour
____ 3.	Dependent variable	c. Factors not studied directly, but could affect the results
____ 4.	Physiological psychology	d. A method for determining the relationship between variables
____ 5.	MacDougall	e. The founder and developer of the school of psycho-analytic psychology
		f. Assumed all behaviour was due to instincts

In this example, the choices are rather cumbersome and need to be read and re-read a number of times before the student can remember and organize them. An improved matching set would be as follows:

Good Item:

	Column A	Column B
____ 1.	The father of British psychology	a. Freud
____ 2.	The founder of Psycho-analytic Psychology	b. Galton
____ 3.	The school of psychology dealing with phenomenology	c. Gestalt
____ 4.	Relegated all behaviour to instincts	d. Jung
____ 5.	Started the first psychology laboratory	e. MacDougall
		f. Skinner
		g. Watson
		h. Wundt

True-False Items

Depending upon the approach involved, there can be quite a variety of basic principles for the construction of true-false items. For the sake of clarity, as well as convenience, nine basic principles for the development of true-false items will be examined.

The basic principles of true-false item construction are:

(1) True-false items are based on statements that are absolutely true or false. There are no exceptions or qualifications involved.

(2) Great care should be taken in avoiding double-loaded statements that are partially true and partially false.

(3) Ambiguous and loosely worded statements are to be avoided.

(4) The main point of a true-false statement should be in a readily noticeable position.

(5) Nonessential material, that could confuse the student should be omitted.

(6) Long and involved statements with a number of qualifying phrases should be avoided.

(7) Double negative statements should never be used and negative statements should not be used if at all possible.

(8) Trick statements should not be used as true-false items.

(9) Indicate the word or words that must be changed when employing the modified true-false item.

Now to cover each of these basic principles in detail.

1. True-false items are based on statements that are abolutely true or false. No exceptions or qualifications are to be included in the item.

The main idea behind this principle is to eliminate the confusion that results when the student is not sure of what the instructor wants. In this situation, the problem that the student encounters is determining the degree or

level of truthfulness of the statement, or the depth of analysis that the instructor expects. For example:

Poor Item:

T F The method of recognition is used to determine whether learning has occurred.

This statement is generally true and would probably be accepted by most students at face value. However, the superior student might look at this item and get the idea that this is a very profound item organized to indicate whether they understand the difference between learning and retention. If the student is aware that the method of recognition is one of the ways of assessing retention of learned material, he is likely to mark the statement as being false.

The instructor's task, therefore, is to run a critical review of each true-false item to determine whether qualifications or exceptions have been inadvertently included in the items. For example:

Good Item:

T F The method of recognition is used to determine the amount of retention of learned material.

2. Great care should be taken in avoiding double-loaded statements that are partially true and partially false.

Occasionally the main point of the true-false item is overlooked by the student when several different ideas are included in the item. For example:

Poor Item:

T F B. F. Skinner has written numerous articles and books about operant behaviour such as The Behaviour of Organisms and The Interpretation of Dreams.

In this item, the student may answer the item in a particular way because he is only familiar with a portion of the item content. In the item above, the

student may well answer the item as true because he is aware that B. F. Skinner has written many articles and books about operant behaviour. The item, however, would have to be false as Sigmund Freud was the author of The Interpretation of Dreams.

The apparent rationale behind the development of the double or multi-loaded statements was the belief that a true-false item could be made more comprehensive in nature by adding several different ideas. Unfortunately, adding the different ideas tends to produce more measurement problems due to the confusion encountered by the students.

By splitting a double or multi-loaded item into its component parts and using each part as a complete item, a valid measurement can be obtained, as well as a wider range in terms of sampling course content. Taking the last example, we would have:

Good Item:

T F B. F. Skinner has written numerous articles and books about operant behaviour.

or

T F The Behaviour of Organisms was written by B. F. Skinner

or

T F B. F. Skinner is the author of The Interpretation of Dreams

Here, we have taken each of the separate themes and made them separate items which reduces the confusion and allows for a more accurate assessment of student behaviour.

3. Ambiguous and loosely worded statements are to be avoided.

As is often the case, an ambiguous true-false item can be interpreted in different ways. The student who encounters an ambiguous item is forced to

spend a fair amount of time trying to decide which meaning the instructor had in mind when the item was constructed. The real difficulty is that if the student guesses the wrong meaning, he may well be penalized, when he may have known the correct answer had he chosen another meaning. For example:

Poor Item:

T F Maturation of behaviour depends upon changes in the nervous system.

The intended answer is true, on the basis that the nervous system is involved in the coordination and organization of the various organ systems. However, the way in which the item is phrased, the student may interpret it as the only thing involved in maturation of behaviour is changes in the nervous system, which is false, as the muscular, endocrine, and circulatory systems are also involved in the maturation process. As a result, the instructor must be on the lookout for ambiguities, even in the very simple type of true-false items.

Another apparently straight-forward true-false item, that upon closer inspection is quite ambiguous, is as follows:

Poor Item:

T F Psychology supports the doctrine of individual differences

The intended answer to this item is true, on the basis of such things as mental and physical characteristics. However, if the student is operating on the basis of the premise that all men are created equal, the answer could then be false. The consequences of ambiguity is to confuse the student and hence, limit the value of the test.

4. The main point of a true-false statement should be in a readily noticeable position.

True-false items should be set in such a way that the main point of the statement will receive the student's attention. It is not uncommon for instructors to construct true-false items that try to disguise the main point from

the student. The result is that the student may miss the item due to misreading, even though he may have the understanding of the material that the item is supposed to measure. The idea is that by the instructor trying to disguise the main point, the purpose of the item is defeated. For example:

Poor Item:

T F If the WAIS was administered, as an experiment, the standardized instructions would be designated the dependent variable.

In this example, the main point - "as an experiment" - appears to be tossed into the item as an aside. The intended answer is false, as the standard instructions would be the independent variable.

The item can be improved by emphasizing the main point. For example:

Good Item:

T F Considering the administration of the WAIS as an experiment, the standardized instructions would be the independent variable.

5. Non-essential material, that could confuse the student, should be omitted.

The common term applied when non-essential material is placed in an item is "eye-wash." All too often, instructors add the non-essential information with the idea that it serves to further clarify or develop interest for the student. While clarification and development of interest have merit, they are of value only if they do not interfere with the process of measurement. The major difficulty with the "eye-wash" is that the instructor tends to go "overboard" and the achievement test becomes a test of reading comprehension and reading speed.

The common problem with the inclusion of non-essential material is that all too often it disguises the main point of the item, thus reducing the validity. For example:

Poor Item:

T F Intelligent industrial personnel workers, workers who have a great deal of ability without much formal education and training, can carry out the administration and scoring of group tests utilized in the industrial setting, providing they have supervision by qualified people.

In this sample item, a number of students would have difficulty separating the "eye-wash" from the main point being evaluated. The item can be improved quite readily by eliminating the non-essential material. For example:

Good Item:

T F Intelligent industrial personnel workers without specialized training can administer and score group tests utilized in the industrial setting.

A common occurrence in the use of "eye-wash" is to construct an item, where the "eye-wash" is the major source of confusion and ambiguity. For example:

Poor Item:

T F Every student taking a course in developmental psychology has many misconceptions to correct.

With the "eye-wash" this item is of no value, as the instructor would not be able to mark the item in a valid manner until it is determined that every student has a number of misconceptions to correct.

6. Long and involved statements with a number of qualifying phrases should be avoided.

The greatest problem that is encountered when long, involved statements make up the item, is that the student is unable to focus on the main point. For example

Poor Item:

T F Regarding learning ability, psychologists would predict that, if a five-year-old were able to suspend all psychological experience for ten years, but was to continue to develop physiologically, and then was tested for learning ability, he would learn faster than he did at the age of five.

If we are to take the position that a classroom test, developed by the instructor, is to assess the students' achievemt in a particular area, we should eliminate as many sources of contamination as we possibly can. For instance, superfluous material in the item statements tends to make the test a measure of reading ability or perhaps a measure of general intelligence rather than a test of achievement. The superfluous material in the last example can be eliminated quite easily, thus making the central theme readily recognizable. A better approach to measuring understanding is as follows:

Good Item:

T F Psychologists would predict that if all psychological experience of a five-year-old were suspended for a ten year period, but physiological development continued in a normal manner, the child would learn faster than at age five.

7. Double negative statements should never be used and negative statements as such, should not be used if at all possible.

Occasionally, there is a need for negative type true-false items, such as when you are trying to assess student understanding of negative concepts. For instance:

Good Item:

T F The definition of learning says nothing about improvement in behaviour.

or

T F Spanking is not the only answer to child misbehaviour.

When negative style items are employed, it is a good idea to focus attention to the negative term by underlining it.

While the negative-type items do serve a purpose, widespread use of this style leads to some serious difficulties such as: confusing the student, complicating or distorting the intended meaning, and the reduction of validity of the test by careless errors that occur when the student overlooks the negative term.

Perhaps the greatest difficulty involved in the negative true-false items is that most of the items involve some very strange and exotic reasoning. That is, in this type of item, you are asking the student to determine that the answer is false, as it is not true that the statement is not true. If you feel confused regarding that explanation, imagine how the student must feel when he tries to work his way through the item.

It is felt that the instructor can accomplish a great deal more by utilizing the direct approach as to whether the statement is true or false. For example:

Poor Item:

T F The object of developmental psychology is not to explain the mental life of children.

We can test this same concept more readily by making the statement so that it can be directly answered true or false.

Good Item:

T F The object of developmental psychology is to describe the mental life of children.

or

T F The object of developmental psychology is to describe and explain sequences of change in children.

8. Trick statements should not be used as true-false items.

The major difficulty with the trick item is that all too often it measures something quite different than what the instructor had intended. Likely as not, the instructor as well as the students, do not have a clue as to what that something really is. For instance, the trick item test could measure intelligence, vigilance, or distrustfulness, rather than assessing progress toward a specific educational objective. In addition to not knowing what is being measured, the trick item test can have some very detrimental effects on students, such as: development of resentful feelings and development of negative attitudes toward tests in general. The net result is a decrease in test validity. For example:

Poor Item:

T F The chief criterion of a good theory is that it guides further research.

The intended response is true, as a good theory does guide further research in the field, however, the student who has some understanding of theory development will answer false, because the major criterion of a good

theory is to allow the experimenter to predict. Thus, many students would get the item wrong due to confusing the terms guiding and predicting.

9. Indicate the word or words that must be changed when employing the modified true-false item.

In an attempt to determine whether students really have an understanding of a concept, many instructors have switched over to using the modified true-false item. The main operation involved in the modified true-false item is that when the item is false, the task of the student is to change the statement so that it becomes true. The most common problem to evolve is that the instructor does not provide any clues as to which words in the statement need to be changed. Another problem, closely related, is that omitting clues may produce a marking nightmare. For example:

Poor Item:

T F Wilhelm Wundt and E. Bradford Titchner developed the Behavioural School of Psychology.

On an item such as this, some students will substitute Structural for Behavioural school. Others will substitute the name of John B. Watson for Wundt and Titchner. Now, while this may be a minor difficulty in marking this particular item, on the more complex items it could mean a very long and laborious job of marking.

To avoid marking problems, underline the word Behaviourist and instruct the students to substitute for the underlined word if they think the statement is false. For example:

Good Item:

T F Koffka and Wertheimer were instrumental in founding the Functional school of psychology.

Completion Items

The basic principles in the development of completion items are as follows:

(1) External clues to the correct response should be avoided.

(2) Answers should be limited to one or two specific words or phrases to avoid ambiguity.

(3) As a generalization, if recall is not being assessed, the completion-type item is not used.

(4) Usually, no more than two completions should be made in any single item.

(5) The blank should be placed at the end of the completion statement, or as near to the end as possible.

(6) The degree of accuracy expected should be specified when using completion items.

Now for a closer inspection of each of these principles.

1. External clues to the correct response should be avoided.

It is quite common to find instructors inadvertently providing clues for the student by the grammatical structuring of the item. For example:

Poor Item:

The definition of learning says nothing about an _____ in behaviour.

In this item, the use of the term "an" would indicate to the alert student that answers such as permanance, practice, and change are ruled out as they would constitute improper English. Hence, it would be fairly easy for the alert student to fill in the correct response - improvement. By making a minor modification to the item structure, the item can be improved immensely.

Good Item:

When defining learning, no mention is made of _____.

One other fairly common clue is provided when the instructor uses a short line to indicate a short answer and a long line to indicate a long word or answer. This minor problem can be overcome very easily by making all of the lines of the same length.

2. Answers should be limited to one or two specific words or phrases to avoid ambiguity.

The instructor should screen each completion item to assure that it calls for a specific word or phrase. Unfortunately completion items are often so loosely framed that the student does not understand what is expected, or worse, is able to evade the intended material under question.

Poor Item:

The central theme in Freud's Psychoanalytic theory is _____.

With a loosely structured item of this type, the instructor should not be surprised at the fantastic variety of student responses. For instance, therapy, developmental theory, motivational theory or personality theory might be common responses. In addition, the terms libido, id, ego, or superego might also be found fairly regularly.

A better way to structure the item to reduce the possibility of evasion or misinterpretation is:

Good Item:

Psychosexual development was a central theme in the personality theory of _____.

or

To eliminate the artificiality produced by the experimental method we use the technique of _____.

In these samples, we have provided direction to the student by structuring the item to reduce the possibility of misinterpretation or evasion.

3. As a generalization, if recall is not being assessed, the completion-type item is not used.

The main purpose for using the completion-type of item is that you are assessing the student's ability to recall the material rather than to provide clues as in the recognition approach. One problem encountered quite often is that the instructor assembles an item that outwardly resembles a completion item, but the structure is in effect, a recognition-type item. For example:

Poor Item:

A child with a C.A. of 6 and an M.A. of 9, has an IQ _____ average.

There are three possible answers to this problem: above, below, or at the average. The net result is that recall *per se* is not being evaluated, and you have in effect, a multiple-choice or true-false type of item. With a bit of luck and a limited amount of knowledge, the student can guess the correct response to this type of item. A much better structure for the item would be:

Good Item:

A child with a C.A. of 6 and an M.A. of 9, has an IQ of _____.

With this structure, recall of the process for calculating IQ is measured. In addition, the scoring procedure becomes quite objective and ambiguity is eliminated.

4. Usually, no more than two completions should be made in any single item.

Many instructors are of the opinion that the more completions in an item, the greater the test of ability to recall the material. The main difficulty with this line of reasoning is that when too many blanks are encountered in an item, scoring problems increase very rapidly and the item becomes very time-consuming for the student. For example:

Poor Item:

The _____ method in child study requires the _____ to _____ conditions which may be _____ to the _____ variable.

The method under assessment is the experimental method as it relates to child study. However, with the five blanks in this item, the student has greater leeway, rather than the rigid structuring that was originally intended. The scoring procedure would be a nightmare for the instructor and the loose structure could lead to a great deal of hassle from the students.

5. The blank should be placed at the end of the completion statement, or as near to the end as possible.

The major reason for placing the blank at or near the end of the statement is efficiency. That is, when the blank is at the beginning of the statement, the student has to read through the statement first, then go back through the statement to the beginning to determine what word or phrase he will use in the blank. The student with limited ability may get lost, especially if the statement is rather lengthy. For example:

Poor Item:

_____ are statistics that describe the test performance of specified groups.

Good Item:

Statistics that describe test performance of specified groups are called _____.

The item is improved immensely when it is shifted around to a straightforward approach where the blank is at the end of the statement. Providing the student knows the material at all, he will be prepared to write in the correct response (norms) by the time he finishes reading the statement.

6. The degree of accuracy expected should be specified when using completion items.

This principle is primarily concerned with completion items that are designed to assess arithmetic, mathematical, or statistical knowledge. Hopefully, the instructor is attempting to evaluate the ability of the student to understand and apply concepts. However, the student will not be able to answer an item correctly if he does not know what is expected. For instance:

Poor Item:

Under normal conditions, the body temperature of the human being is _____.

Most students would write in the answer of 37 and should receive credit for it. However, the correct response should be 37 C. As a result, a better approach, that offers better direction and organization would be:

Good Item:

Under normal conditions, the body temperature of the human being is _____ ^{o}C.

Essay Items

To produce greater understanding and clarity, the principles for development of essay-type items will be listed, then each will be covered separately.

(1) To improve sampling, use a number of short essay items, rather than concentrating on one or two long essay items.

(2) The purpose of essay items is to assess goals that are difficult to assess with the other types of items.

(3) The degree of detail expected should be indicated in each essay item.

(4) The instructor using essay items should set definite limits as to the scope of each item.

Taking each principle separately, we have:

1. To improve sampling, use a number of short essay items rather than concentrating on one or two long essay items.

The major reason for employing a number of short essay items rather than one or two long essay items, is that it allows the instructor a much broader sampling of the learnings in the course. This in turn, increases the reliability and validity of the assessment of the students' achievement.

2. The purpose of essay items is to assess goals that are difficult to assess with the other types of items.

A major factor resulting in less efficient ability to test the objectives for specific course content is the failure of the instructor to restrict the scope of the essay items. A common tendency among instructors is the use of the essay item for gathering general information. Generally, assessing information can be achieved more effectively and efficiently by using objective items. Consider the following item:

Poor Item:

Name and locate the major structural components of the human nervous system.

This type of essay item ignores the major purposes of the essay item, such as problem-solving, analysis, or making judgments, and concentrates on the gathering of general information. In addition, this type of essay item can produce a variety of difficulties related to marking, as well as restricting the opportunities for written expression. The following essay item by contrast, indicates the special merits of the essay approach to assessment.

Good Item:

"If identical twins are sep rated early in life, one going to an average educational environment and the other to an intellectually impoverished environment, the twin in the impoverished environment would be handicapped in all learning." Do you agree or disagree with this statement? Why?

On an essay item structured in this manner, the student must evaluate the relevant information and determine the best way of expressing the response before writing. Thus, it appears that the essay item, when properly structured, evaluates different mental abilities than those utilized in answering objective items.

3. The degree of detail expected should be indicated in each essay item.

If the purpose of an essay item is to determine the extent of the student's ability to analyze information when no specific directions are provided, the following item will suffice:

Poor Item:

"Explain how psychology differs from the physical sciences."

The instructor using this item may accept two differences as an adequate answer to this item. When the students are unaware of this scoring

arrangement, they may discuss a single difference between psychology and the physical sciences or they may attempt to discuss a half dozen possible differences.

When the instructor's marking system only provides for full credit for discussing two of the differences the students should be informed, as in the following example:

Good Item:

"Explain two ways in which psychology differs from the physical sciences."

Another of the problems related to the degree of detail in the essay item is the selection of qualifying words. That is, what the instructor wants the student to do, such as: "describe," "differentiate," "list," "explain", or "state." A frequent occurrence is for the instructor to use these terms as synonyms, when each demands a specific type of student behaviour.

4. The instructor using essay items should set definite limits as to the scope of each item.

A very common problem with the essay item is that the student is supposed to write everything he knows about a very general topic. While freedom to respond is a basic aspect of the essay item, instructors should set the limits in which the student is free to respond. For example:

Poor Item

"Describe the development of psychology from the establishment of the first experimental laboratory to the present time."

The first problem that comes to mind on a loosely structured item like this, is how is the instructor to set up a rating or marking system? With no restrictions or defined limits, one would expect a wide variety in student responses to this item.

By limiting the amount of material expected from the student, we will have a better chance of obtaining a uniform sample of student understanding. For example:

Good Item:

"Write a paragraph on the topic of comparative psychology."

If any degree of comparison is to be achieved with essay items, the students must be provided with directions about a specific situation. For example:

Good Item:

Show how each of the following has been a factor in the development of the field of psychology:

a) The Structural School
b) The Functional School
c) The Behaviourism School

This item will provide at least a minimum level of comparison for marking, as well as providing a frame of reference for the students.

SECTION IX

THE ORGANIZATIONAL COMPONENTS OF CLASSROOM TESTS

The organizational components of a test refer to such things as how the test is arranged, how it is reproduced, instructions to the students and how the test will be marked. The common tendency is to spend a great deal of time on the construction of test items, after which little or no care is taken with the organization of the items into a test. The organizational components are of value for both the students and the instructor.

Format of the Items

The first generalization is that each item should be presented in its entirety on the same page. Here, it is often times the typist who produces the problem. That is, half of the item appears on one page and the other half on the following page, or a graph is on one page and references to the graph are on other pages. When items are scattered onto two or more pages, the students are exposed to unnecessary as well as unappreciated problems. Usually a short seminar with the typist will resolve the issue.

A second format component is the numbering of the items. Numbering the items may not be of much value while the student is taking the test, but will be of great value in identifying specific items for discussion after they are marked.

For the true-false, matching and completion items, the arrangement of the items on the page is quite standard. However, in dealing with the multiple-choice item, several different approaches have been used. Probably the best

item arrangement from the students' viewpoint is the following, where each of the choices is distinct.

Good Item Arrangement

The part of a camera to which the iris of the eye corresponds is the:

a) diaphragm
b) lens
c) film plate
d) bellows

While this item arrangement offers the greatest clarity, often times the amount of space may be a prime consideration. As a result, when the choices are short, space may be conserved by the following arrangement:

Fair Item Arrangement

A test which requires the examinee to respond in a precise manner to the items is said to be:

a) Structured c) Stable
b) Standardized d) Subjective

Hopefully, space will not be the major consideration in the arrangement of items, so the following may be avoided:

Poor Item Arrangement

In educational and psychological testing, objectivity refers to the: (a) administration of the test, (b) format of the test items, (c) scoring procedures used, (d) method of interpreting scores.

Arrangement and Ordering of Items

The next problem that confronts most instructors is whether to group the items by type or by the content they are intended to measure. That is, should all matching items be grouped together, or should items dealing with specific course content be grouped together regardless of the type of item?

The generalization is that items that are similar in type should be grouped together. However, for an objective examination that is over 100 items in length, fatigue effects can be reduced by changing the item style several

times. For example, the first 25 items might be multiple-choice, then 10 items of the matching variety, then 20 completion style items, followed by another 20 multiple-choice, and so on.

Another area of controversy is concerned with whether the items should be arranged in the sequence in which they were taught, or assembled in terms of level of difficulty. Ideally, the test should be constructed as a power-type test, that is, begin with easy items and add more and more difficult items as the test becomes longer. In this manner, all students should be able to answer some of the items. The problem with the power-type test is that most instructors do not keep a record of the difficulty level of their test items. Hence, as the instructor looks at the item, it may appear easy to him, but prove to be extremely difficult for the students. In spite of the drawbacks in determining the level of difficulty for the items, it appears to be worthwhile for the instructor to set the test in a power-type framework.

It is possible, with a little extra effort, to set the test as several subtests, so that each subtest is arranged as a power test, yet each subtest covers only a single area of content. This, of course, calls for special directions to the students to continue with the test in spite of running into difficult items.

Correct Response Distribution

As there is a chance factor operating in the responses to objective items, it is of value to minimize the degree of guessing correct responses rather than knowing the material. As a result, it is good practice to mix the items so that five or six will not all be "b" or all "d" responses. In addition, we have found that some instructors tend to use a patterned sequence for the multiple-choice correct responses, such as: abcd or dcba. The students may not know much of the material, but you better believe that they will look for a pattern or sequence in the items they know for sure.

With true-false items, a Gellerman series will generally suffice for reducing chance or guessing correct responses. That is, the correct responses would be in this sequence: T F F T T F T F F T

Scoring Procedures

To save hours in flipping over pages to score items which are answered on the test booklet, we feel it is worthwhile to design a special answer sheet. The answer sheet can be scored quickly by making up a key or template that reveals the correct responses in a glance. In addition to saving time in scoring, the second advantage of the separate answer sheet is that the test booklets may be collected after going over the items and thus, the items may be used again on a different form of the test.

As most instructors tend to number the items on their test, it is of value to use letters to indicate the choices for each item. This practice avoids confusion between the item number and the choice number. Usually multiple-choice items do not have over five choices. As a result, the answer sheet for a mixed multiple-choice, true-false, and completion test might be as follows:

1. a b c d e
2. a b c d e
3. a b c d e
4. a b c d e
5. a b c d e
6. T F
7. T F
8. T F
9. T F
10. T F
11. _____
12. _____ etc.

If a marking template is to be used, the student is directed to mark an X through the choice they consider to be correct, or to print the correct response for the completion items on the appropriate line. In this manner, marking 50-items for each of 35 students can be accomplished in about one hour.

One last factor related to scoring procedures is the allotment of credit. If each item on the test is worth one point, this should be stated at the beginning of the test. If, however, the multiple-choice are one point each and the matching items are worth two points each, this fact should be made clear to the students.

Probably the greatest source of complaints related to the allotment of credit is the essay or short answer type item. On these items, the amount of credit allotted usually indicates the depth or extent of the response expected by the instructor. For example:

Good Scoring Procedure

1. (10 points) Describe the circumstances under which each of two "schools of psychology" has appeared in the development of the field of psychology.

2. (5 points) Show how one of the "schools of psychology" had an important effect on the direction of present day psychology.

Test Directions

Regardless of how much exposure students have had with various types of tests, the best bet for the instructor is to assume that students are not familiar with tests, especially of the objective variety. With this orientation, the instructor tends to take greater pains to make sure the test directions are clear and concise and in a language that can be understood by all of the students. Further, it is of value to go through the directions orally with the students, as they have a common tendency to either not read thoroughly or to ignore the directions. To test for this lack of attention, use the following statement as directions on the answer sheet without going through the directions orally:

> Place an X through the letter which indicates your choice for each item, e.g. a X c d e

Then, pass out the test booklet and answer sheets and have them go to work. Perhaps, from their earlier experiences, anywhere from 25% to 50% of the class will circle the letter indicating their choice. Most of the difficulties regarding directions will be reduced if you read through the directions with the class.

Should We Correct for Guessing

The last aspect related to the organizational components of classroom tests deals with the pros and cons of correcting for guessing. While there are a variety of viewpoints on the use of correction formulas for guessing, we feel that the procedure is not worth the cost in time and effort for the classroom situation.

Our purpose in this booklet has been to attempt to help instructors to make their tests more meaningful, without getting carried away and forgetting the task at hand. The same approach applies with correction for guessing formulas. In fact, instructors should encourage their students to attempt to answer all of the items. While many instructors feel that this approach promotes guessing, it can be utilized as an excellent teaching technique. That is, when instructors go over each item with their students, they should not only cover the correct response, but explain why the other choices are not correct. This procedure helps to make students aware of the importance of reading each of the choices very carefully.

Comments About Pretesting

Apparently over the years, instructors at various levels of education have come to believe that they receive in their classrooms or lecture halls, a group of empty heads, and they are charged with the task of filling these heads with all of the educational goodies available in their particular field of endeavour. Where we got off on this viewpoint no one seems to know, however, it would appear to be time to get out of this rut and take a look at what we are doing. The point, unless you are working with some very handicapped students, it would appear to be quite reasonable that students coming into our classes are going to have some knowledge of the material that we are attempting to have them learn. If we take the viewpoint that students come into class with some amount of background, then it behooves us to find out how much they know prior to beginning the course material.

The major reason for doing a pretest for a particular course is to obtain an indication of the level of knowledge for each student and then assess his or her progress from that level. At present, we tend to assume that all students begin at the zero level and then make progress from that point. Many instructors would be quite surprised that the students who receive the top marks in their classes would have received high marks on their tests, even without the benefit of instruction. If our marking system is to indicate the amount or quality of learning, perhaps the idea of pretesting deserves a closer inspection. For example, we administer a pretest to a particular class. We find that George has a pretest score of 72%, and that Pierre has a pretest score of 28%. Then, as the course material is covered and we administer the final examination, we find that George is at 81%, but Pierre is at the 74% level. Who has learned more in this situation? Should we be thinking of the pretest as a means for getting at the individualization of instruction and learning that we have been talking about for years?

SECTION X

BIBLIOGRAPHY

American Psychological Association, Standards for Educational and Psychological Tests and Measurements. Washington, D.C.: American Psychological Association, 1966.

American Psychological Association, Technical Recommendations for Psychological Tests and Manuals, Washington, D.C., 1966.

Anastasi, A., Differential Psychology, 3rd ed., New York: Crowell-Collier & Macmillan, 1960.

Anastasi, A. (Ed.), Testing Problem in Perspective. Washington, D.C.: American Council on Education, 1966.

Anastasi, A., Psychological Testing, 3rd ed., New York: Crowell-Collier & Macmillan, 1968.

Barnette, W.L. Jr. (Ed.), Readings in Psychological Tests and Measurements. Nobleton, Ontario: Irwin-Dorsey Ltd., 1968.

Beggs, D.L. & Lewis, E.L., Measurement And Evaluation In The Schools. New York: Houghton-Mifflin, 1975.

Buros, O., (Ed.), Mental Measures Yearbooks V-VII. Highland Park, New Jersey: Gryphon Press, 1959-1972.

Buros, O., (Ed.), Tests in Print, Highland Park, New Jersey: Gryphon Press, 1974.

Buros, O., (Ed.), Tests in Print II, An Index to Tests, Test Reviews and the Literature on Specific Tests, Highland Park, New Jersey: Gryphon Press, 1974.

Cattell, R.B., Abilities: Their Structure, Growth, and Action, Boston, Mass.: Houghton-Mifflin, 1971.

Chase, C.I., Measurement For Educational Evaluation, Reading, Mass.: Addison-Wesley Publishing Co., 1974.

Cronbach, L., Essentials of Psychological Testing, 3rd ed., New York: Harper & Row, 1970.

Cronbach, L.J. & Drenth, P.J.D., (Eds.), Mental Tests and Cultural Adaptation, The Hague, Netherlands: Mountan Publishers, 1972.

Davis, F.B., Educational Measurements and Their Interpretation, Belmont, California: Wadsworth, 1964.

Ebel, R.L., Measuring Educational Achievement, Englewood Cliffs, New Jersey: Prentice-Hall, 1965.

Educational Testing Service, Multiple-Choice Questions: A Close Look, Princeton, New Jersey: Educational Testing Service, 1963.

Goldman, L., Using Tests in Counselling, New York: Appleton-Century-Crofts, 1961.

Goslin, D., The Search for Ability: Standardized Testing in Social Perspective, New York: Russell Sage Foundation, 1963.

Goslin, D.A., Teachers and Testing, New York: Russell Sage Foundation, 1967.

Gronlund, N.E., Measurement And Evaluation in Teaching, 3rd ed., New York, N.Y.: Macmillan Publishing Co., 1976.

Guilford, J.P. & Hoepfner, R., The Analysis of Intelligence, New York: McGraw-Hill, 1971.

Helmstadter, G., Principles of Psychological Measurement, New York: Appleton-Century-Crofts, 1964.

Hills, J.R., Measurement and Evaluation in Schools, Columbus, Ohio: Merrill Publishing Co., 1976.

Holt, R.R. (Ed.), Diagnostic Psychological Testing, New York: International University Press, 1968.

Horst, P., Psychological Measurement and Prediction, Belmont, California: Wadsworth, 1966.

Journal of Applied Psychology, Washington, D.C.: American Psychological Association, 1917 - present.

Journal of Counselling Psychology, Washington, D.C.: American Psychological Association, 1954 - present.

Journal of Educational Measurement, East Lansing, Michigan: National Council on Measurement In Education, 1964 - present.

Karmel, L.J., Measurement And Evaluation In The School, New York, N.Y.: Macmillan Publishing Co., 1970.

Lien, A., Measurement And Evaluation Of Learning, 3rd ed., Dubuque, Iowa: W.C. Brown, 1976.

Lyman, H., Test Scores and What They Mean, Englewood Cliffs, New Jersey: Prentice-Hall, 1963.

Magnusson, D., Test Theory, Reading, Mass.: Addison-Wesley, 1966.

Mehrens, W.A. & Lehmann, I.J., Measurement and Evaluation in Education and Psychology, New York, New York: Holt, Rinehart & Winston, 1973.

Nelson, C.H., Measurement And Evaluation In The Classroom, New York, N.Y.: Macmillan Publishing Co., 1970.

Personnel Psychology, Durham, North Carolina: Personnel Psychology, Inc., 1948 - present.

Rapaport, D., Gill, M. & Schafer, R., Diagnostic Psychological Testing, (Revised edition by R. Holf, Ed.), New York: International Universities Press, 1968.

Rosenthal, R. & Jacobsen, L., Pygmalion In the Classroom: Teacher Expectation and Pupils' Intellectual Development, New York: Holt, Rinehart & Winston, 1968.

Storey, A., Measurement Of Classroom Learning: Teacher Directed Assessment, Palo Alto, California: Science Research Associates, 1970.

Test Service Notebook Series, New York, N.Y.: Harcourt, Brace and World Inc., yearly publications.

Thorndike, R. (Ed.), Educational Measurement, rev. ed., Washington, D.C.: American Council on Education, 1969.

Thorndike, R.L. & Hagen, E.P., Measurement And Evaluation In Psychology and Education, 3rd ed., New York, New York: Wiley and Sons, 1969.

Tyler, L.E., Tests and Measurements, 2nd ed., Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

SECTION XI

APPENDICES

Appendix A Test Publishers Directory

Appendix B Selected Tests of General Aptitude

Appendix C Selected Tests of General Achievement

Appendix D Selected Tests of Reading Achievement

APPENDIX A

Test Publisher's Directory (* Canadian)

American Guidance Service, Inc.
Publisher's Building, Circle Pines, Minnesota 55014

Bobbs-Merrill Co., Inc.
4300 West 62nd Street, Indianapolis, Indiana 46268

Bruce Publishers
340 Oxford Road, New Rochelle, New York 10804

CTB/McGraw-Hill
Del Monte Research Park, Monterey, California 93940

Consulting Psychologists Press, Inc.
577 College Avenue, Palo Alto, California 94306

Educational And Industrial Testing Service
P.O. Box 7234, San Diego, California 92107

Educational Testing Service
Princeton, New Jersey 08540

Follett Educational Corporation
1010 W. Washington Blvd., Chicago, Illinois 60607

* Guidance Centre
College of Education, University of Toronto
1000 Yonge Street, Toronto, Ontario N4W 2K8

Guidance Testing Associates
6516 Shirley Avenue, Austin, Texas 78752

* Harcourt Brace Jovanovich Inc.
(Academic Press Canada)
55 Barber Greene Road, Don Mills, Ontario M3C 2A1

Houghton-Mifflin Co.
110 Tremont Street, Boston, Massachusetts 02107

Institute for Personality and Ability Testing
1602 Coranado Drive, Champaign, Illinois 61820

* Institute of Psychological Research, Inc.
34 Fleury Street West, Montreal, Quebec H3L 1S9

Klamath Printing Company
320 Lowell Street, Klamath Falls, Oregon 97601

* McGraw-Hill Ryerson Ltd.
330 Progress Avenue, Scarborough, Ontario M1P 2Z5

* Nelson & Sons Canada Ltd.
81 Curlew Drive, Don Mills, Ontario M3A 2R1

Psychological Corporation
304 East 45th Street, New York, N.Y. 10017

Psychologists and Educators Press
419 Pendik, Jacksonville, Illinois 62650

Psychometric Affiliates
Box 3167, Munster, Indiana 46321

Public Personnel Association
1313 East 60th Street, Chicago, Illinois 60637

Scholastic Testing Services, Inc.
480 Meyer Road, Bensenville, Illinois 60106

Science Research Associates, Inc.
259 East Erie Street, Chicago, Illinois 60611

Teachers College Press
502 West 121st Street, New York, New York 10027

Western Psychological Services
12031 Wilshire Blvd., Los Angeles, California 90025

Xerox Publishing Company
275 Wyman Street, Waltham, Massachusetts 02154

APPENDIX B

Selected Tests of General Aptitude*

Test and Publisher	Grade Levels	Types of Scores Obtained
A. Single-Score Tests		
Henmon-Nelson Test of Mental Ability (Houghton-Mifflin)	3-College Grad.	
Kuhlman-Anderson Measure of Academic Potential (Psychological Corporation)	K-13	
Otis-Lennon Mental Ability Test (Harcourt, Jovanovich, Inc.)	K-13	
B. Two-Score Tests		
Academic Promise Test (Psychological Corporation)	6-9	Verbal, Nonverbal, & Total
California Tests of Mental Maturity (CTB/McGraw-Hill)	K-College & Adult	Language, Nonlanguage & Total
Lorge-Thorndike Intelligence Tests (Houghton-Mifflin)	K-13	Verbal, Nonverbal
School and College Ability Tests (Educational Testing Service)	4-14	Verbal, Mathematical, & Total

Test and Publisher	Grade Levels	Types of Scores Obtained
C. Multifactor Tests:		
Differential Aptitude Tests (Psychological Corporation)	8-12 & Adult	Verbal Reasoning, Numerical Ability, Abstract Reasoning, Space Relations, Mechanical Reasoning, Clerical Speed & Accuracy, & Language Usage
Flanagon Aptitude Classification Tests (Science Research Associates)	9-12	Reasoning, Vocabulary, Judgment & Comprehension
SRA Primary Mental Abilities (Science Research Associates)	K-12	Verbal meaning, perception, number, space, and reasoning
Tests of Educational Ability (Science Research Associates)	4-12	Language skills, reasoning skills, and Quantitative skills

APPENDIX C

Selected Tests of General Achievement*

Test and Publisher	Grade Levels	Types of Areas Covered
A. Content-Oriented Tests		
Canadian Achievement Tests (CTB/McGraw-Hill Ryerson Ltd.)	1.6 - 12.9	Reading: Vocabulary, Comprehension, Spelling. Language: Mechanics, Expression. Mathematics: Computation, Concepts, Applications, Reference skills (Norm- & Criterion-Referenced)
Comprehensive Tests of Basic Skills (CTB/McGraw-Hill)	2 1/2 - 12	Reading, Language, Arithmetic and Study Skills
Metropolitan Achievement Tests (Harcourt Brace Jovanovich, Inc.)	1 - 12	Vary with level, such as advanced: Work Knowledge, Reading, Spelling, Language, Arithmetic, Social Studies and Sciences
SRA Achievement Series (Science Research Associates)	2 - 9	Reading, Arithmetic, Language and Study Skills
Stanford Achievement Test (Harcourt Brace Jovanovich, Inc.)	2 - 9	Reading, Arithmetic, Language, Social Studies, Science, and Study Skills
B. Application Tests		
Canadian Tests of Basic Skills (Thomas Nelson & Sons Canada, Ltd.)	3 - 12	Understanding Basic Social Concepts, Background in Natural Sciences, Correctness and Appropriateness of Expression, Ability to: do Quantitative Thinking, Interpret Material in: Social Studies, Natural Sciences, Library Material, etc. (Measures what the student can do, not what he knows)

APPENDIX D

Selected Tests of Reading Achievement*

Test and Publisher	Grade Levels	Types of Scores Obtained
Botel Reading Inventory (Follett Educational Corp.)	1 - 12	Frustration, Instructional, and Free Reading Scores
Durrell Analysis of Reading Difficulty (Harcourt Brace Jovanovich, Inc.)	1 - 6	Oral Reading, Silent Reading, Listening Comprehension, Word Analysis, and Recognition, Spelling, Handwriting
Gates-MacGinite Reading Tests (Teachers College Press)	1 - 12	Vocabulary, Speed and Accuracy, and Comprehension
Diagnostic Reading Scales ((Spache -) McGraw-Hill)	1 - 7	Word Recognition, Word Analysis, Comprehension and Phonics
Standard Reading Inventory (Klamath Printing Company)	1 - 7	Independent, Minimum, Maximum and Frustrating Reading Levels - Also - Vocabulary, Word Recognition, Comprehension and Speed Scores
McCullough Word Analysis Tests (Xerox Publishing Company)	4 - 6	Initial Blends, Digraphs, Phonetic Discrimination, Matching Letters to Vowel Sounds, Sounding Words, Interpreting Phonetic Symbols, Syllabication, Finding Root Words
Gates-McKillop Reading Diagnostic Tests (Teachers College Press)	1 - 7	Vocabulary, Oral Reading, Phrases, Syllabication, Auditory Discrimination, Spelling, Blending, Vowels
Diagnostic Reading Test (Scholastic Testing Service, Inc.)	2 - 6	Scores in Single Consonants, Vowel Sounds, Silent "E", Syllabication

Test and Publisher	Grade Levels	Types of Scores Obtained
Durrell-Sullivan Reading Capacity and Achievement Tests (Harcourt Brace Jovanovich, Inc.)	2 - 6	Vocabulary and Comprehension
Iowa Silent Reading Tests (Harcourt Brace Jovanovich, Inc.)	4 - 12	Rate, Comprehension, Word Meaning and Locating Information
Kelley-Greene Reading Comprehension Test (Harcourt Brace Jovanovich, Inc.)	9 - 12	Comprehension, Directed Reading, Retention
Nelson-Denny Reading Test (Houghton-Mifflin)	9 - 12	Vocabulary, and Paragraph Comprehension
SRA Reading Record (Science Research Associates)	8 - 12	Rate, Comprehension, and Vocabulary

* All of these tests as well as all other types of tests may be purchased through:

The Guidance Centre
371 Bloor Street West
Toronto, Ontario

OR

The Institute of Psychological Research
34 Fleury Street West
Montreal, Quebec

GLOSSARY

Academic or Scholastic Aptitude. The combination of native and acquired abilities that is needed for academic or school work. Scores indicate the likelihood of success in mastering academic work, based on estimates from measures of the necessary abilities.

Achievement Age. The age for which a given achievement test score is the estimated average. That is, if the achievement age corresponding to a score of 28 on a reading test is 9 years, 2 months (9-2), this means that students 9 years, 2 months of age achieve, on the average, a score of 28 on that test.

Achievement Test. A test that measures the extent to which the individual has acquired certain skills or information, usually as a result of instruction.

Age Norms. Values representing typical or average performance for individuals in various age groups.

Alternate-Form Reliability. The correlation between results on alternate (or parallel) forms of a test. A measure of the extent to which the two forms of the test measure the same content consistently or reliably.

Aptitude. The combination of native and acquired characteristics and abilities that appear to indicate an individual's ability to learn. The learning could be general or specific in nature. A specific aptitude might be "mechanical aptitude" which refers to a combination of mental abilities, motivational factors, as well as interests, which are conducive to acquiring proficiency in the mechanical field.

Average. The general term applied to measure of central tendency. The most widely used measures of central tendency are the mean, median, and mode.

Battery. Generally, a battery refers to a group of standardized tests administered to the same group of people, so the results on the tests are compared. Occasionally, the term battery is applied to any group of tests administered together, even though they are not standardized on the same subjects.

Ceiling. The upper limit of ability measured by a particular test.

Coefficient of Correlation (r). A measure of the degree of relationship between two sets of measures for the same group of individuals. Values of (r) run from +1.00, a perfect positive correlation to -1.00, a perfect negative correlation. A value of (r) = 0.00 indicates no relationship between the two measures.

Completion Item. A test question calling for the addition of a word, phrase, or sentence to complete the statement.

Correction for Guessing. Here, the score is lowered for wrong answers. Generally this technique is applied to multiple choice or true-false tests. While the technique is questioned by many people in the field, it is still employed on some tests. The most common approach is right minus wrong and the result is the individual's score.

Correlation. The relationship between two sets of scores on the same group of people. The degree and direction of the relationship are denoted by the coefficient of correlation (r).

Criterion. The standard by which a test is evaluated. The criterion is usually another set of scores or ratings that the test scores are designed to predict or are correlated with.

Diagnostic Test. A test used to locate and determine the extent of areas of strength or weakness. Most common are diagnostic achievement tests in the areas of reading, spelling and arithmetic.

Discriminating Power. The degree to which a test item is able to differentiate between persons having a great deal of some trait versus those possessing very little of that trait.

Equivalent Form. Any of two or more forms of a test that are closely related in terms of the content being measured and the difficulty of the items.

Forced-Choice Item. Refers to any multiple-choice item where the subject is required to select one or more of the given choices.

Grade Norm. The average score obtained by students in a given grade.

Group Test. A test that may be administered to any number of students at the same time. Usually, only one administrator is needed, however, several proctors should be on duty to assist and observe for unusual conditions.

Individual Test. A test that can only be administered to one person at a time.

Inventory Test. Tests that are designed to cover in detail, some relatively small area of instruction or training. Many personality and interest inventories appraise the individual's status on personality characteristics or on a wide variety of activities.

Item. A single question or exercise on a test.

Matching Item. A test item in which the individual must correctly associate the response words in one column with the stimulus words in another column.

Mean. The arithmetic mean, obtained by adding all the scores in a distribution and dividing by the number of scores in the distribution.

Median. The middle score in a distribution of scores that are ranked from highest to lowest.

Mode. The most frequently occurring score in a distribution.

Multiple-Choice Item. A test item made up of a stem statement and usually four or five choices, where the subject's task is to select the correct response from those available.

Multiple-Response Item. A special type of multiple-choice item, where two or more responses may be correct.

N. The symbol denoting the number of scores or people making up the distribution.

Norms. The statistics that describe the test performance of a particular group, usually of a specific age or grade level.

Objective Test. A test where the judges would all agree as to which response was correct. For example, multiple-choice, matching, or true-false items. The objective test may be contrasted with the "subjective" test, where each judge might assign a different value to each item, such as the essay examinations.

Performance Test. A test requiring a manual or motor response by the subject, often involving the manipulation of objects. A performance test may also be an actual work sample to determine whether the subject can handle a particular task, such as the ability to type.

Personality Test. A test used to measure non-intellectual aspects of an individual's behaviour, such as personality organization or personality adjustment.

Power Test. A test designed to measure the level of performance or ability rather than the speed of response. Usually power tests do not have a time limit.

Practice Effect. Where previous experience with the test carries over and influences the subject's performance on the same or similar tests. The shorter the period between testings, the greater the practice effect.

Profile. Generally, the graphic presentation of the results of several tests for the same individual. Allows for fast, simple identification of strengths and weaknesses as measured by test performance.

Prognostic Test. A test primarily used to predict future performance in a field or in the work setting.

Raw Score. The numerical result obtained when scoring a test. It is usually the number of correct responses on a test and has no meaning other than the size of the number until it is compared with some standard, such as the performance of the whole group on a particular test.

Readiness Test. A test measuring the degree of maturity or acquired skills needed for going on to profit from the next level of learning.

Recall Item. A question that requires the subject to develop a correct response from memory with no clues as to the correct response. For example, the essay item: Describe Piaget's contributions to psychology.

Recognition Item. A question or statement which requires the subject to recognize the correct response from a number of given responses. The best example is the multiple-choice item.

Standardized Test. A test that is administered, scored, and interpreted according to pre-determined conditions that is given to a a representative group of people of a particular age or grade level. The test results for these people serve as a reference point for interpreting the results of other people taking the same test.

Subjective Test. A test in which each judge assigns the marks on the basis of his own interpretation. For example, having five English instructors independently mark the same essay.

Survey Test. A test indicating a subject's general achievement in a particular field or academic area. The survey test is usually concerned with the status of the group in a subject area, rather than with individuals within the group.

True-False Item. A question is presented and the subject has to determine whether it is true or false.

ERIC Clearinghouse for Junior Colleges JUL 12 1991